Testing for parameter stability in nonlinear autoregressive models*

Claudia Kirch† Joseph Tadjuidje Kamgaing‡

July 8, 2011

Abstract

In this paper we develop testing procedures for the detection of structural changes in nonlinear autoregressive processes. For the detection procedure we model 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, can be extended to this case. The limit distribution under the null hypothesis is obtained, which is needed to construct asymptotic tests. For a large class of alternatives it is shown that the tests have asymptotic power one. In this case, we obtain a consistent change-point estimator which is related to the test statistics.

Power and size are further investigated in a small simulation study with a particular emphasis on situations where the model is misspecified, i.e. the data is not generated by a neural network but some other regression function. As illustration, an application on the Nile data set as well as S&P log-returns is given.

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

AMS Subject Classification 2000: 62G10, 62M45,62G08

1 Introduction

The question of structural stability of models is very important in diverse areas of science such as economy, finance, hydrology, physics or quality control. In statistics the field of

*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

change-point analysis has a long tradition dating back to Page [42, 43], who introduced it in the context of quality control. It deals with the question whether the stochastic structure of an observed time series has changed at some unknown point in the sample. For a detailed overview of the field of change-point analysis we refer to the book by Csörgő and Horváth [12].

During the past decades change-point problems have been attracting more and more interest and have been investigated in different ways. A classical scenario is a possible mean change in otherwise independent identically distributed random variables. Later on, this was extended to stability tests of the parameters of a regression function. Most of these tests are based on the variational or dynamical behavior of the partial sum process $\hat{S}_n(k)$ of the estimated residuals frequently by considering weighted maxima of the partial sum process. In this context, CUSUM based tests introduced by Brown et al. [8] are of particular importance. For a change in the mean and a change in the parameter of a linear model CUSUM-based tests have been investigated to a great extend (confer Csörgő and Horváth [12] and some of the references therein). These tests have then been extended to linear autoregressive time series by several authors. For example, Kulperger [34] as well as Horváth [21] use test statistics based on partial sums of residuals for linear autoregressive time series. Similary, Bai [4] makes use of partial sums of residuals for the change analysis in ARMA-models. Such tests based on estimated residuals have the advantage of being easily calculable and having a good power for those alternatives they can detect. On the other hand, they usually only detect changes of the unconditional mean. Davis et al. [13] obtain a Gaussian-type likelihood ratio statistic which asymptotically detects all parameter changes. However, least-squares estimators need to be calculated for each possible change-point. Hušková et al. [25] propose a related approach using weighted sums of residuals, which overcomes this difficulty.

For the nonlinear setting most of the change-point literature deals with jumps in the regression function using a Kernel based approach and not with distributional changes over time (confer Müller [39] or Delgado and Hidalgo [14]). Little is done for nonlinear time series models, an exception is found in Andrews [3], where Wald, Lagrange Multiplier and Likelihood ratio tests based on generalized method of moments are designed for testing parameter stability and structural changes. As for linear autoregressive models these tests require the calculation of estimators for any possible change-point.

In this paper we consider nonlinear autoregressive time series, which have recently been attracting some attention in the time series literature (confer e.g. the textbook by Tong [51] and references therein). White [52] as well as Teräsvirta et al. [50] compare the forecasting performance of a linear model with those of nonlinear models including models where the regression function is given by neural networks. Several other authors including Francq et al. [17] and Luukkonen et al. [35] propose statistical tests to distinguish between linear and non-linear autoregressive time series.

In this paper we develop change-point tests for nonlinear time series based on estimated residuals, where we approximate the nonlinear regression function by a neural network to estimate these residuals. The asymptotic theory is then derived for general nonlinear regression functions. Due to its universal approximation property (confer e.g. White [53]

or Franke et al. [18]) a large class of functions can be approximated by a neural network to any degree of accuracy. 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. This is also confirmed by our small simulation study where a special emphasis is given to autoregressive time series where the regression function is not a neural network (cf. Section 4). Approximations via different parametric models than neural networks can be derived in a similar manner under appropriate assumptions on the parametric regression function.

We consider the nonlinear autoregressive time series

$$X_t = g(\mathbb{X}_{t-1}) + \varepsilon_t, \tag{1.1}$$

where $\mathbb{X}_{t-1} = (X_{t-1}, \ldots, X_{t-p})$ and $\{\varepsilon_t : 1 \leqslant t \leqslant n\}$ are independent identically distributed random errors having a positive variance and satisfying further conditions specified below.

The model defined in equation (1.1) covers a wide class of processes including the single layer feedforward neural network based autoregression as in (1.4) below, the classical linear autoregressive (AR) processes as well as the threshold autoregressive (TAR) models due to Tong [51]. The existence of strictly stationary solutions of equation (1.1) as well as their geometric ergodic property can be derived using the stability theory for Markov chain, see e.g. Meyn and Tweedie [37] or Tong [51]. An and Huang [1] give further conditions that can easily be checked and are applicable to a large range of processes including the ones mentioned above. Further, the independence assumption of the residuals to the past observations of the process considered in An and Huang [1] appears to be standard in the literature and induces the causality of the strict stationary solution of equation (1.1).

In view of possible financial applications, an indirect example of model (1.1) is given by the nonlinear ARCH processes, see for example Stockis et al. [48], an extension of the β-ARCH model introduced by Guégan and Diebolt [20] which include the celebrated ARCH model defined by Engle [16]. Applying the logarithm to the squared nonlinear ARCH time series transforms it into a model following (1.1). Back to the change point problem, Berkes et al. [22, 5, 6, 7] developed several tests for both a change in a parameter of the GARCH process as well as changes in the residual distribution or correlation structure for the important example of ARCH and GARCH-processes.

As mentioned above, we will develop change-point tests based on estimated residuals, which we obtain by approximating the nonparametric function g by a one layer feedforward neural network with H hidden neurons

$$f(\mathbf{x}, \theta) = \nu_0 + \sum_{h=1}^{H} \nu_h \psi(< \boldsymbol{\alpha}_h, \mathbf{x} > + \beta_h), \tag{1.2}$$

where $<,>$ is the classical scalar product on $\mathbb{R}^p$ and $\boldsymbol{\alpha}_j = (\alpha_{j1}, \ldots, \alpha_{jp})$. In this paper 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

$$\lim_{x \to -\infty} \psi(x) = 0, \qquad \lim_{x \to \infty} \psi(x) = 1, \qquad \psi(x) + \psi(-x) = 1. \tag{1.3}$$

A popular choice is the logistic function

$$\psi(x) = (1 + e^{-x})^{-1}.$$

Denote $\theta = (\nu_0, \ldots, \nu_H, \boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_H, \beta_1, \ldots, \beta_H)$. As a special case of (1.1) we obtain an autoregressive time series model with a neural network as regression function

$$X_t = f\left(\mathbb{X}_{t-1}, \theta_0\right) + \varepsilon_t \tag{1.4}$$

for some θ_0. We will use this as an approximation to the true model (1.1) to obtain estimated residuals. This is a reasonable approach due to the universal approximation property of neural networks (White [53] or Franke et al. [18]). Time series following this model will be called correctly specified, otherwise we will call them misspecified.

Stockis et al. [48] use the time series in (1.4) as building blocks in a regime-switching model, the so called CHARME-models, in the context of financial time series. In their model the duration time in each regime is random and driven by a hidden Markov chain, while in classical change-point analysis the duration time is usually fixed and deterministic. For CHARME-models they use the theory of stability of Markov chains as developped in Meyn and Tweedie [37] or Tong [51] to prove the existence of a unique strictly stationary and causal solution of the CHARME-model. Since Model (1.4) is a special case with only one regime, the following result is a straightforward application of Theorem 4 of Stockis et al. [48].

Theorem 1.1. *Let ε_t have a density function that is positive on the real line and θ_0 belong to a compact set. Then, X_t has a unique strictly stationary and causal solution, which is geometrically ergodic. Furthermore, if there exists $m > 0$ such that $\mathbb{E}|\varepsilon_t|^m < \infty$, then, $\mathbb{E}|X_t|^m < \infty$.*

Based on this model we can derive a least-squares estimator $\widehat{\theta}_n$ of the parameter θ_0 and consequently can estimate the residuals by $\widehat{\varepsilon}_t = X_t - f(\mathbb{X}_{t-1}, \widehat{\theta}_n)$. Under misspecification this leads to an approximation of the true regression function by a neural network. Precisely we minimize the nonlinear least squares (NLLS) with respect to θ

$$Q_n(\theta) \;:=\; \sum_{t=p+1}^{n} \left(X_t - f(\mathbb{X}_{t-1}, \theta)\right)^2 =: \sum_{t=p+1}^{n} q_t(\theta),$$

thus we consider the nonlinear least squares estimator

$$\widehat{\theta}_n = \arg\min_{\theta \in K} Q_n(\theta) \tag{1.5}$$

for a suitable compact set K. The minimization is usually obtained by solving the nonlinear score function

$$\frac{\partial Q_n(\widehat{\theta}_n)}{\partial \theta} = 0,$$

which yields

$$\sum_{t=p+1}^{n} \widehat{\varepsilon}_t = 0. \tag{1.6}$$

Since we do not assume that $\{X_t\}$ is a neural-networked based autoregressive time series as in (1.4), the question arises how this estimator behaves under misspecification (1.1). This is investigated in detail in Section 2.

Description of the test procedure

Let us now consider a time series model with a change after an unknown time point $1 \leqslant k^* = k^*(n) \leqslant n$

$$X_t = \begin{cases} X_t^{(1)}, & t \leqslant k^* \\ X_t^{(2)}, & t > k^*, \end{cases} \tag{1.7}$$

where $\{X_t^{(1)}\}$ as well as $\{X_t^{(2)}\}$ are time series that differ distributionally. While this is not needed to obtain the below theoretic results, we have in mind that $\{X_t^{(1)}\}$ as well as $\{X_t^{(2)}\}$ are both of the form (1.1) but with different regressions functions. In the correctly specified model both follow (1.4) but with different parameters. The unknown parameter k^* is called the change-point if $k^* < n$. For $k^* = n$ no change occurs.

We are now interested in the testing problem

$$H_0 : k^* = n \qquad \text{vs.} \qquad H_1 : k^* < n.$$

Our testing procedures are based on various functionals of the partial sums of estimated residuals using the least-squares estimator $\widehat{\theta}_n$ as in (1.5)

$$\hat{S}_n(k) = \sum_{t=p+1}^{k} \widehat{\varepsilon}_t = \sum_{t=p+1}^{k} \left(X_t - f(\mathbb{X}_{t-1}, \widehat{\theta}_n) \right).$$

In order to investigate the behavior of statistics based on $\hat{S}_n(k)$, we need to understand how the estimator $\widehat{\theta}_n$ behaves in case of the misspecified model (1.1) as well as under the alternative (1.7) with $k^* < n$, which is also considered in Section 2.

The minimization procedure takes place with respect to some suitable compact set K. If $\widehat{\theta}_n$ is not in the interior of this compact set, we will reject the null hypothesis immediately since either a change occurred or model (1.4) is not capable of describing the observed

time series sufficiently well. Otherwise one of the following test statistics is used:

$$
\begin{aligned}
T_{n1} &= \max_{p<k<n} \left(\sqrt{\frac{n-p}{k(n-p-k)}}\, |\hat{S}_n(k)| \right), \\
T_{n2}(q) &= \max_{p<k<n} \left(\frac{1}{\sqrt{n-p}\, q(\frac{k}{n-p})}\, |\hat{S}_n(k)| \right), \\
T_{n3}(G) &= \max_{p+G<k\leqslant n} \frac{1}{\sqrt{G}} \left| \hat{S}_n(k) - \hat{S}_n(k-G) \right|, \\
\widetilde{T}_{n3}(G) &= \max_{p+G<k\leqslant n-p-G} \frac{1}{\sqrt{2G}} \left| \hat{S}_n(k+G) - 2\hat{S}_n(k) + \hat{S}_n(k-G) \right|, \\
T_{n4}(r) &= \frac{1}{n-p} \sum_{k=p+1}^{n-1} \frac{1}{r(k/(n-p))} \left(\frac{1}{\sqrt{n-p}} \hat{S}_n(k) \right)^2,
\end{aligned}
\tag{1.8}
$$

where $q(\cdot)$ and $r(\cdot)$ are weight functions defined on $(0,1)$ specified below and $G < n$.

Theorem 3.1 gives the null asymptotics for the above statistics, from which we can derive an asymptotic size α test. Theorem 3.2 gives some assumptions under which these tests have asymptotic power one in the correct as well as misspecified model. In this situation Corollary 3.1 shows how to obtain a consistent estimator for the change-point.

In some applications data is observed sequentially and a decision whether a change has occurred or not has to be made online. An example are financial time series such a stock returns where adjustments of investment strategies should be made if a change occurred. The procedures discussed in this paper can also be extended to this situation, for details we refer to Kirch and Tadjuidje-Kamgaing [29].

The remainder of the paper is organized as follow: Statistical properties of the above NLLS-estimator under the null hypothesis as well as alternatives, the correctly specified as well as misspecified models are discussed in the next section.
Section 3 contains the asymptotic distribution of the proposed test statistics under the null hypothesis as well as some consistency results under alternatives. Results are derived taking possibly misspecified models into account. In addition, a consistent change-point estimator is obtained.
Section 4 illustrates the usefulness of the proposed procedures with a simulation study as well as an application to the Nile data set as well as S&P data.
The final sections contain the proofs.

2 Some Properties of Neural Network Estimators

In order to obtain asymptotics for the change-point statistics, we need to understand the behavior of $\widehat{\theta}_n$ under the null as well as alternative hypothesis, for the correctly as well as misspecified model. For the case of a correctly specified null hypothesis, there is an abundant literature on this topic. Since we present a different proof the moment assumptions required in this case are weaker than for the classical proofs. Using a

uniform law of large numbers for random variables defined on a separable Banach space (cf. Ranga Rao [46]), we need only the second derivative of the cost function, instead of the third as usual for the nonlinear parametric models (see e.g. Theorem 3.2.24 in Taniguchi and Kakizawa [49]). Furthermore, we obtain asymptotic results such as consistency and asymptotic normality with respect to some well-defined parameter $\widetilde{\theta}_0$ for the misspecified model (1.1) as well as in case a change-point is present.

To this end assume

A. 1. Let $\{X_t^{(1)} : t \in \mathbb{Z}\}$, $\{X_t^{(2)} : t \in \mathbb{Z}\}$ be stationary and ergodic processes and $\mathbb{E}|X_1^{(1)}|^\nu < \infty$ and $\mathbb{E}|X_1^{(2)}|^\nu < \infty$ for some $\nu \geqslant 2$.

Under the alternative this is a simplifying assumption to avoid additional technical difficulties. Frequently, we have the following model in mind

$$X_t = \begin{cases} g(\mathbb{X}_{t-1}) + \varepsilon_t, & t \leqslant k^*, \\ h(\mathbb{X}_{t-1}) + \varepsilon_t, & t > k^*. \end{cases} \tag{2.1}$$

In this case the time series after the change-point has starting values from the stationary distribution of the time series before the change-point and is therefore not stationary. The consistency as in Theorem 2.1 remains true in this situation if the uniform law of large numbers holds for the time series after the change. Some results to this end concerned with the law of large numbers as well as central limit theorem can for example be found in Meyn and Tweedie [37], Chapter 17, as well as Jensen and Rahbek [27]. The asymptotic normality of Theorem 2.2 is only needed under the null hypothesis for the purpose of this paper but also remains true for model (2.1) under some additional assumptions. In the simpler situation of a linear autoregressive model Hušková et al. [25] derive an explicit formula for the difference of a time series with starting value $\mathbb{X}_{t-1}^{(1)}$ and starting values from the corresponding stationary distribution. Their results show that for linear time series the starting values are irrelevant in all of the above cases.

Furthermore, we assume:

A. 2. The change-point fulfills $k^* = \lfloor \lambda n \rfloor$ for some $0 < \lambda \leqslant 1$.

Here, $\lambda = 1$ corresponds to the null hypothesis while $0 < \lambda < 1$ corresponds to the alternative hypothesis.

Let

$$E_\theta = \lambda \mathbb{E}(X_t^{(1)} - f(\mathbb{X}_{t-1}^{(1)}, \theta))^2 + (1-\lambda)\mathbb{E}(X_t^{(2)} - f(\mathbb{X}_{t-1}^{(2)}, \theta))^2, \tag{2.2}$$

If no change occurs the expressions simplifies to $E_\theta = \mathbb{E}(X_t - f(\mathbb{X}_{t-1}, \theta))^2$.

Define

$$\widetilde{\theta}_0 = \arg\min_\theta E_\theta \tag{2.3}$$

and assume

A. 3. $\widetilde{\theta}_0$ is the unique minimizer of E_θ and lies in the interior of the compact parameter set $K \subset \mathbb{R}^{H(p+2)+1}$.

The existence of $\widetilde{\theta}_0$ is needed only to obtain the asymptotic distribution of the test statistic under the null hypothesis (compare Theorem 3.1). Under alternatives it is not entirely necessary but simplifies the conditions on what kind of changes are detectable (confer Theorem 3.2).

If the parameter space is chosen in such a way that the network is identifiable, then the true parameter θ_0 in the correctly specified model without change (1.4) is the unique minimizer of $E_\theta = \mathbb{E}(f(\mathbb{X}_{t-1}, \theta_0) - f(\mathbb{X}_{t-1}, \theta))^2$, which shows that $\widetilde{\theta}_0 = \theta_0$.

In order to get a better understanding of identifiability in neural networks, we state the following result of Hwang and Ding [26].

Lemma 2.1. *Assume that*

(i) $f(\mathbf{x}, \widetilde{\theta}_0)$ *is not redundant (i.e. there exists no other networks with fewer hidden neurons* ($H' < H$) *that represent exactly the same relationship function).*

(ii) $f(\mathbf{x}, \widetilde{\theta}_0)$ *is irreducible, i.e., for all* $i \neq 0,\ j \neq 0$,

a) $\nu_i \neq 0$

b) $\alpha_i \neq \mathbf{0}$

c) $(\alpha_i, \beta_i) \neq (\alpha_j, \beta_j)$ *for all* $i \neq j$.

If the activation function satisfies (1.3)*, then up to a family of permutations and transformations defined below,* $f(\mathbf{x}, \widetilde{\theta}_0)$ *is identifiable unique.*

To understand the transformation leading to non-identifiability, redefine $\theta = (\nu_0, \mu_1, \cdots, \mu_H)$ with $\mu_h = (\nu_h, \beta_h, \alpha_h), h = 1, \cdots, H$, then

1. a permutation of $\mu_i = (\nu_i, \beta_i, \alpha_i)$ and $\mu_j = (\nu_j, \beta_j, \alpha_j)$ still provide the same neural networks function, i.e. a permutation of the i-th neuron and j-th neuron will not change the value of the neural networks function.

2. Additionally, by using the relation in equation (1.3) one derives

$$\nu_i \psi(< \alpha_i, x > + \beta_i) = \nu_i (1 - \psi(< -\alpha_i, x > - \beta_i)).$$

   Henceforth, we can easily verify that $(\nu_0, \mu_1, \cdots, \mu_{i-1}, \mu_i, \mu_{i+1}, \cdots, \mu_H)$ and $(\nu_0 + \nu_i, \mu_1, \cdots, \mu_{i-1}, -\mu_i, \mu_{i+1}, \cdots, \mu_H)$ yield the same neural network function.

In practice these transformations do not yield a problem, as in each segment of the parameter space $\widetilde{\theta}_0$ is identifiably unique so that it is still guaranteed that $f(\mathbf{x}, \widehat{\theta}_n)$ is close to $f(\mathbf{x}, \widetilde{\theta}_0)$, which is all that matters for the below change-point test.

Theorem 2.1. *Let A.1 – A.3 hold. Then* $\widehat{\theta}_n$ *is strongly consistent for* $\widetilde{\theta}_0$*, i.e.*

$$\widehat{\theta}_n \to \widetilde{\theta}_0 \qquad a.s. \qquad (n \to \infty).$$

Remark 2.1. In the correctly specified model without change (1.4) the result of Theorem 2.1 can be obtained under the weaker assumption of the existence of only first moments of the innovations by noting that $\widehat{\theta}_n$ is the minimizer of $Q_n(\theta) = \sum_{t=p+1}^{n} \tilde{q}_t(\theta)$ with $q_t(\theta) = (f(\mathbb{X}_{t-1}, \theta) - f(\mathbb{X}_{t-1}, \theta_0))^2 + 2\varepsilon_t(f(\mathbb{X}_{t-1}, \theta) - f(\mathbb{X}_{t-1}, \theta_0))$. The proof is analogous to the one of Theorem 2.1 with the difference that a uniform law of large number is obtained for Q_n under the existence of only first moments of the innovations.

In order to derive asymptotic normality we need some additional assumptions. Recall that a stationary process $\{T_t\}$ is called α- or strong mixing with mixing rate $\alpha(\cdot)$ if

$$\alpha(j) = \sup_{A \in \mathcal{F}^0_{-\infty}(T),\, B \in \mathcal{F}^\infty_j(T)} |P(A \cap B) - P(A)P(B)| \to 0 \text{ as } j \to \infty,$$

where $\mathcal{F}^0_{-\infty}(T)$ is the σ-algebra generated by $T_0, T_{-1}, \cdots$ and $\mathcal{F}^\infty_j(T)$ is the σ-algebra generated by $T_j, T_{j+1}, \cdots$.

A. 4. The time series $\{X_t^{(1)}\}$ and $\{X_t^{(2)}\}$ are independent and α-mixing with $\alpha(j) = O(j^{-c})$ for some $c > \nu/(2-\nu)$, where $\nu > 2$ is such that and $\mathbb{E}|X_0^{(l)}|^\nu < \infty$, $l = 1, 2$.

The mixing assumption is a classical condition in nonlinear time series analysis and follows from stationarity and geometric ergodicity, which can be derived with little effort using the stability theory for Markov processes, see e.g. Meyn and Tweedie [37]. For the correctly specified model 1.4 this follows from a result of Stockis et al. [48] as pointed out in Theorem 1.1. Recently, different dependency concepts have been introduced to tackle some of the deficiencies of mixing conditions (cf. e.g. Doukhan and Louhichi [15] for the weak dependence approach). Our results remain true if these conditions ensure certain assumptions. For example, to obtain the next theorem we need that $\frac{1}{\sqrt{n}}\nabla Q_n(\widetilde{\theta}_0)$ fulfills a central limit theorem, for the results in the next section that it fulfills the law of iterated logarithm in addition to a strong invariance principle for $\zeta_t = X_t - f(\mathbb{X}_{t-1}, \widetilde{\theta}_0)$. The advantage of using mixing conditions is twofold. First, for $\mathbb{X}_t$ mixing, processes of the form $g(\mathbb{X}_t)$ for some measurable g are also strong mixing with the same rate (cf. the proof of Theorem 2.2). Secondly, all the results we need are available in the literature for mixing processes and follow for example from the invariance principle of Kuelbs and Philipp [33].

The independence assumption is a technical condition that is not fulfilled for processes as in (2.1). However, asymptotic independence of the two processes in the sense that (5.1) remains true is sufficient to obtain the result but difficult to prove in the general setting (2.1). For linear autoregressive time series, this follows from an explicit representation of the series with respect to the starting values as has been derived in Hušková et al. [25]. Since we only need asymptotic normality of the estimator under the null hypothesis to derive the asymptotic distribution of the change-point statistics, deriving asymptotic normality under alternatives is only of marginal interest in this paper.

A. 5.

$$A = \nabla^2 E_{\widetilde{\theta}_0}$$

is positive definite, where ∇^2 denotes the Hesse matrix with respect to θ.

Finally, we can state asymptotic normality of the estimators.

Theorem 2.2. *Let A.1 – A.5 hold. Then*

$$\sqrt{n}(\widehat{\theta}_n - \widetilde{\theta}_0) \xrightarrow{\mathcal{L}} \mathcal{N}(0, A^{-1}VA^{-1}),$$

where

$$V = \lim_{n\to\infty} \frac{1}{n}\mathbb{E}\nabla Q_n(\widetilde{\theta}_0)(\nabla Q_n(\widetilde{\theta}_0))^T$$

with $\nabla Q_n(\theta) = \frac{\partial Q_n(\theta)}{\partial \theta}$ *is the gradient of* $Q_n(\theta)$*. The limit* V *exists but may be singular.*

3 Consistency of the Change-Point Tests

In this section we derive the null asymptotics for the test statistics introduced in (1.8). Furthermore, we show that the corresponding tests have asymptotic power one for a large class of important alternatives. In this case, we obtain a consistent estimator for the change-point which is related to the test statistics.

First, we need to introduce some additional notation. We assume that the weight function q belongs to the class

$$Q_{0,1} = \{q \,:\, q \text{ is non-decreasing in a neighborhood of zero, non-increasing in a neighborhood of one and } \inf_{\eta\leqslant t\leqslant 1-\eta} q(t) > 0 \text{ for all } 0 < \eta < 1/2\}.$$

We need additionally that the following integral is finite for at least some $c > 0$

$$I^*(q,c) = \int_0^1 \frac{1}{t(1-t)} \exp\left\{\frac{-cq^2(t)}{t(1-t)}\right\} dt.$$

A very important class of weight functions fulfilling these conditions are

$$q(t) = (t(1-t))^\gamma, \qquad 0 \leqslant \gamma < 1/2,$$

where a γ close to $1/2$ rather detects early or late changes and a γ close to 0 detects changes in the middle. Note that for $\gamma = 1/2$ we obtain statistic T_{n1} (which is asymptotically independent from the statistics with $\gamma < 1/2$).

For details and further references confer Csörgő and Horváth [11], Chapter 4.
We assume that the weight function r fulfills for all $x \in (0,1)$

$$r(x) > 0 \quad \text{and} \quad \int_0^1 \frac{t(1-t)}{r(t)}\, dt < \infty. \tag{3.1}$$

For more details and further references confer Csörgő and Horváth [12], Chapter 2.

Moreover define

$$\alpha(x) = \sqrt{2\log x}, \qquad \beta(x) = 2\log x + \frac{1}{2}\log\log x - \frac{1}{2}\log\pi. \tag{3.2}$$

The key to the next theorem is a strong invariance principle for $\zeta_t = X_t - f(\mathbb{X}_{t-1}, \widetilde{\theta}_0)$, i.e. there exists a Wiener process $\{W(t)\}$ (possibly after enlarging the probability space) and $0 < \kappa < 1/2$, $\tau > 0$ such that

$$\sum_{i=1}^{k}(\zeta_i - \mathbb{E}\zeta_1) - \tau W(k) = O\left(k^{1/2-\kappa}\right) \qquad a.s. \tag{3.3}$$

In case of a correctly specified model (1.4) we get $\zeta_t = \varepsilon_t$ is an i.i.d. sequence, which fulfills the above invariance principle with $\tau^2 = \mathrm{var}(\varepsilon_1)$, $\kappa = (\nu-2)/(2\nu)$. This is a classical result by Komlós et al. [31, 32] and Major [36], which has subsequently been generalized to dependent random variables. For example, Kuelbs and Philipp [33] prove such a limit theorem for strong mixing sequences fulfilling A.4.

The use of mixing conditions has the advantage that the properties immediately carry over to functionals of the time series such as $Y_t = \nabla f(X_t, \widetilde{\theta}_0)$. In principle, the mixing conditions can be relaxed to obtain the below theorem, if the invariance principle (3.3) holds in addition to assumptions on certain functionals of the time series such as $\{Y_t\}$. For example, Wu [54] proves the above invariance principle for certain processes of the type $X_t = g(\varepsilon_t, \varepsilon_{t-1}, \ldots)$, but the results for relevant functionals needed to be derived additionally.

For the statistics $T_{n2}(q)$ and $T_{n4}(r)$ it is sufficient but more technical if only a functional central limit theorem in addition to some Hájek-Rényi-type inequalities (for $q, r \neq 1$) holds (for some technical details we refer to Kirch [28], proof of Corollary 6.1).

We are now ready to state the asymptotic distribution under H_0 for the above statistics.

Theorem 3.1. *Let the null hypothesis of no change hold, i.e. $\lambda = 1$. Furthermore assume A.1 – A.5.*

a) Then we have for all $x \in \mathbb{R}$

$$P\left(\alpha(\log(n-p))\frac{T_{n1}}{\tau} - \beta(\log(n-p)) \leqslant x\right) \longrightarrow \exp(-2e^{-x}) \qquad as\ n \to \infty.$$

b) If $q \in Q_{0,1}$ and $I^*(q,c) < \infty$ for some $c > 0$, then

$$\frac{1}{\tau}\, T_{n2}(q) \xrightarrow{\mathcal{D}} \sup_{0<t<1} \frac{|B(t)|}{q(t)} \qquad as\ n \to \infty.$$

c) If $G = G(n) \to \infty$, $\frac{G}{n} \to 0$ and $G^{-1}n^{1-2\kappa}\log n \to 0$ as $n \to \infty$, κ as in (3.3), then we have for all $x \in \mathbb{R}$

$$P\left(\alpha((n-p)/G)\frac{T_{n3}(G)}{\tau} - \beta((n-p)/G) \leqslant x\right) \longrightarrow \exp(-2e^{-x}) \quad as\ n \to \infty.$$

d) If $G = G(n) \to \infty$, $\frac{G}{n} \to 0$ *and* $G^{-1}n^{1-2\kappa}\log n \to 0$ *as* $n \to \infty$, κ *as in* (3.3), *then we have for all* $x \in \mathbb{R}$

$$P\left(\alpha((n-p)/G)\frac{\widetilde{T}_{n3}(G)}{\tau} - \beta((n-p)/G) + \log(2/3) \leqslant x\right) \longrightarrow \exp(-2e^{-x}) \quad \text{as } n \to \infty.$$

e) If r *fulfills condition (3.1), then*

$$\frac{1}{\tau^2}\, T_{n4}(r) \xrightarrow{\mathcal{D}} \int_0^1 \frac{B^2(t)}{r(t)}\, dt \quad \text{as } n \to \infty.$$

Here $\{B(t) : 0 \leqslant t \leqslant 1\}$ *denotes a Brownian bridge. The assertions remain true if instead of the true long-run standard deviation* $\tau > 0$ *an estimator* $\widehat{\tau}_n$ *is used as long as for b) and e)* $\widehat{\tau} - \tau = o_P(1)$, *for a)* $\widehat{\tau} - \tau = o_P\left((\log\log n)^{-1}\right)$ *and for c) and d)* $\widehat{\tau} - \tau = o_P\left((\log n/G)^{-1}\right)$.

The following lemma gives a variance estimator as is needed for the change-point tests above in case of a correctly specified model. It can still be reasonably used in applications if the approximation is good (cf. Remark 3.1).

Lemma 3.1. *Under A.1 – A.5 it holds for the correctly specified model* (1.4) *under the null hypothesis*

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

if $2 \leqslant \nu < 4$ *in A.1. If* $\nu \geqslant 4$, *then we get the stronger rate* $O_P(n^{-1/2})$.

Remark 3.1. In practical applications it is advisable to use the following adapted variance estimator, which – under appropriate assumptions – is also a consistent estimator under alternatives in the fully correctly specified situation, where $\{X_t^{(2)}\}$ follows (1.4) with a different parameter but an innovation-sequence with the same variance as before the change-point:

$$\begin{aligned}\widehat{\sigma}_{a,n}^2 &= \frac{\widehat{k^*}}{n}\frac{1}{\widehat{k^*} - (H(p+2)+1)} \sum_{t=p+1}^{\widehat{k^*}} \hat{\varepsilon}_t^2 \\ &\quad + \left(1 - \frac{\widehat{k^*}}{n}\right)\frac{1}{n - \widehat{k^*} - (H(p+2)+1)} \sum_{t=\widehat{k^*}+1}^{n} \hat{\varepsilon}_t^2 \end{aligned} \tag{3.4}$$

where $\widehat{k^*}$ is as in Corollary 3.1

$$\hat{\varepsilon}_t = \begin{cases} X_t - f(\mathbb{X}_{t-1}, \hat{\theta}_1), & t \leqslant \widehat{k^*}, \\ X_t - f(\mathbb{X}_{t-1}, \hat{\theta}_2), & t > \widehat{k^*}, \end{cases}$$

$$\hat{\theta}_1 = \arg\min_{\theta} \sum_{t=p+1}^{\widehat{k^*}} \left(X_t - f(\mathbb{X}_{t-1}, \theta)\right)^2, \hat{\theta}_2 = \arg\min_{\theta} \sum_{t=\widehat{k^*}+1}^{n} \left(X_t - f(\mathbb{X}_{t-1}, \theta)\right)^2.$$

Theorem 3.1 suggests to use an estimator for the long-run variance τ^2 instead of σ^2 in order to obtain asymptotically the correct size even under misspecification. To this end we propose to use a flattop-kernel estimator taking possible changes into account with the automatic bandwidth selection procedure by Politis [44]

$$\widehat{\tau}^2 = \widehat{\tau}^2(\Lambda_n) = \max\left(\widehat{R}(0) + 2\sum_{k=1}^{\Lambda_n} w(k/\Lambda_n)\widehat{R}(k), \frac{1}{n}\widehat{\sigma}^2_{a,n}\right), \tag{3.5}$$

where for $a > b$ we define $\sum_a^b = 0$, $\widehat{R}(k) = \frac{1}{n}\sum_{t=1}^{n-k}\hat{\varepsilon}(t)\hat{\varepsilon}(t+k)$ and

$$w(t) = \begin{cases} 1, & |t| \leqslant 1/2, \\ 2(1-|t|), & 1/2 < t < 1, \\ 0, & t \geqslant 1. \end{cases}$$

The maximum on the right-hand side of the formula is needed to guarantee that the estimator is positive but remains scale invariant. In simulations we use the following automatic bandwidth selection procedure proposed by Politis [44]:
Automatic bandwidth selection procedure: Let $\widehat{\lambda}$ be the smallest positive integer such that $\left|\widehat{R}(\widehat{\lambda}+k)/\widehat{R}(0)\right| < 1.4\sqrt{\log_{10} n/n}$, for $k = 1, \ldots, 3$. Then choose the bandwidth $\widehat{\Lambda}_n = 2\widehat{\lambda}$.
Hušková and Kirch [23] discuss the equivalent of this estimator in case of mean changes in dependent data.

The problem is that the long-run variance τ^2 is much more difficult to estimate than σ^2 so that we do not obtain such good estimates for small sample sizes (cf. e.g. the simulation study in Hušková and Kirch [23]). On the other hand if the approximation of X_t by a neural network based autoregressive process with parameter $\widetilde{\theta}_0$ is good enough, the two variances are almost equal, but the variance estimator (3.4) is much more accurate than the long-run variance estimator (3.5). If – for small and medium samples sizes – the estimation error of (3.5) is larger than the approximation error involved when using (3.4), then the latter one yields the better results. This is confirmed by the simulation study in Section 4, where change-point tests based on the variance instead of long-run variance estimator lead to a better performance if measure by empirical size and power.

Theorem 3.1 enables us to construct tests based on the above statistics with the correct asymptotic size. In the next theorem it is shown that those tests have asymptotic power one under a large class of alternatives.

Before we can state the theorem we first need to give some additional assumptions.

A. 6. (a) There exists $c > 0$ such that

$$P\left(\max[|\mathbb{E}X_1^{(1)} - \mathbb{E}f(\mathbb{X}_p^{(1)}, \theta)|_{\theta=\widehat{\theta}_n}|\,, |\mathbb{E}X_1^{(2)} - \mathbb{E}f(\mathbb{X}_p^{(2)}, \theta)|_{\theta=\widehat{\theta}_n}|]) \geqslant c\right) \to 1.$$

(b) There exists $c > 0$ such that

$$P\left(\left|(\mathbb{E}X_1^{(1)} - \mathbb{E}f(\mathbb{X}_p^{(1)}, \theta)|_{\theta=\widehat{\theta}_n}) - (\mathbb{E}X_1^{(2)} - \mathbb{E}f(\mathbb{X}_p^{(2)}, \theta)|_{\theta=\widehat{\theta}_n})\right| \geqslant c\right) \to 1.$$

Remark 3.2. If the assumptions of Theorem 2.1 hold, Assumption A.6 a) simplifies to

$$|\mathbb{E}f(\mathbb{X}_p^{(1)}, \widetilde{\theta}_0) - \mathbb{E}X_1^{(1)}| > 0.$$

In this situation by the definition of $\widetilde{\theta}_0$ and A.3

$$|\mathbb{E}f(\mathbb{X}_p^{(1)}, \widetilde{\theta}_0) - \mathbb{E}X_1^{(1)}| = \frac{1-\lambda}{\lambda}|\mathbb{E}X_1^{(2)} - \mathbb{E}f(\mathbb{X}_p^{(2)}, \theta)|.$$

If $\widetilde{\theta}_0$ does not exist under alternatives, the behavior of $\widehat{\theta}_n$ is arbitrary, but a) is still fulfilled if

$$\min_{\theta \in K}(\max(|\mathbb{E}f(\mathbb{X}_p^{(1)}, \theta) - \mathbb{E}X_1^{(1)}|, |\mathbb{E}X_1^{(2)} - \mathbb{E}f(\mathbb{X}_p^{(2)}, \theta)|)) > 0. \tag{3.6}$$

Analogous expressions can be obtained for b). Equation (3.6) essentially means that there exists no neural network $f(\mathbf{x}, \theta)$ with H hidden neurons for which $\mathbb{E}f(\mathbb{X}_p^{(1)}, \theta) = \mathbb{E}X_1^{(1)}$ as well as $\mathbb{E}f(\mathbb{X}_1^{(2)}, \theta) = \mathbb{E}X_1^{(2)}$. In the simple mean-change-model without neural networks (i.e. trivial neural networks where $H = 0$) both assumptions reduce to $\mathbb{E}X_1^{(1)} \neq \mathbb{E}X_1^{(2)}$.

Assumptions like these are typical for change-point statistics that are based on estimated residuals and even occur in a simple linear regression situation, where tests based on estimated residuals only detect changes if the unconditional expectation changes (cf. e.g. Hušková and Koubkova [24]). Tests, which can detect general alternatives in the correctly specified model, can usually be obtained by using partial sums of vector-weighted estimated residuals, yet they are theoretically and computationally much more complicated. In a nonlinear parametric setting such as in (1.4) tests with power against all type of changes in the typically high-dimensional parameter space will by construction be less powerful than tests looking for specific important alternatives. The approach discussed in this paper only detects a restricted class of important alternatives but does so very successfully as the simulation study in Section 4 shows. Additionally, it is computationally and theoretically easier accessible than tests based on high-dimensional weighted sums of residuals.

In Kirch and Tadjuidje-Kamgaing [30] we show that the tests have asymptotic power one even under certain local alternatives. The key tool is a uniform central limit theorem, which replaces the uniform law of large numbers (Theorem 5.1) in the below proof.

Theorem 3.2. *Assume that A.1 holds, where it is sufficient if the moment assumptions holds for $\nu = 1$, and the minimization takes place in a compact set K. Furthermore, the change-point fulfills Assumption A.2 with $0 < \lambda < 1$.*

a) Let Assumption A.6 (a) hold.

(i) For all $c \in \mathbb{R}$ we get

$$P\left(\alpha(\log(n-p))T_{n1} - \beta(\log(n-p)) \geqslant c\right) \to 1.$$

(ii) If $q \in Q_{0,1}$*, then*

$$T_{n2}(q) \xrightarrow{P} \infty,$$

which means that $P(T_{n2}(q) \geqslant c) \to 1$ *for all* $c > 0$*.*

(iii) If $G = G(n) \to \infty$*,* $\frac{\log n}{G} \to 0$ *but* $G/n \to 0$*, then we have for all* $c \in \mathbb{R}$

$$P\left(\alpha((n-p)/G)T_{n3}(G) - \beta((n-p)/G) \geqslant c\right) \to 1.$$

(iv) If r *fulfills condition (3.1), then*

$$T_{n4}(r) \xrightarrow{P} \infty.$$

b) Let Assumption A.6 (b) hold. If $G = G(n) \to \infty$*,* $\frac{\log n}{G} \to 0$ *but* $G/n \to 0$*, then we have for all* $c \in \mathbb{R}$

$$P\left(\alpha((n-p)/G)\widetilde{T}_{n3}(G) - \beta((n-p)/G) + \log(2/3) \geqslant c\right) \to 1.$$

We obtain asymptotic power one in case of an unknown variance, if the variance estimator is at least stochastically bounded under the alternative.

Based on the partial sum process $\{\hat{S}_n(k)\}$ we additionally obtain a consistent estimator for the change-point as the following corollary shows.

Corollary 3.1. *Let Assumptions A.1 (*$\nu \geqslant 1$ *is sufficient), A.2 with* $0 < \lambda < 1$ *as well as A.3 and A.6 (a) hold. Then*

$$\frac{\widehat{k^*}}{n} \xrightarrow{P} \lambda,$$

where

$$\widehat{k^*} = \arg\max\left\{\left|\hat{S}_n(k)\right| : 1 \leqslant k < n\right\}. \tag{3.7}$$

4 Simulation Study and Real Data Applications

In the previous sections we have shown that the derived tests have asymptotic level α and power one for a large class of alternatives.

In Section 4.1, we consider the behavior of the statistic if the data really is an autoregressive process generated by a neural network. As in reality this will generally not be fulfilled, the most pressing question is what happens under misspecification, which will be considered in Section 4.2. In this context the question arises how sensitive the procedure is with respect to the choice of the number of hidden neurons which is also

considered there. Finally in Section 4.4, we would like to have a more detailed look at what kind of alternatives can be detected (cf. also Remark 3.2).

For illustrational purposes the methods will then be applied to two data sets which have been used frequently in the context of change-point detection, namely the Nile data set as well as S&P-returns in Section 4.5.

Because of limitations of space we restrict the simulations to the statistic $T_n := T_{n2}(q)$ with $q \equiv 1$, changes at $k^* = n/2$ and an autoregression of order 1. For the implementation of the test optimization algorithms to obtain the parameter $\widehat{\theta}_n$ as in (1.5) (resp. for $\hat{\theta}_1, \hat{\theta}_2$ as in Remark 3.1) are needed. To this end the mathlab algorithm *fminsearch* is used and the data set only included in the simulations if the variable *exitflag* of *fminsearch* indicates that the algorithm has converged. The idea behind this is that the test would only be applied in situations were this is the case and the approximation by a neural network based process is reasonable.

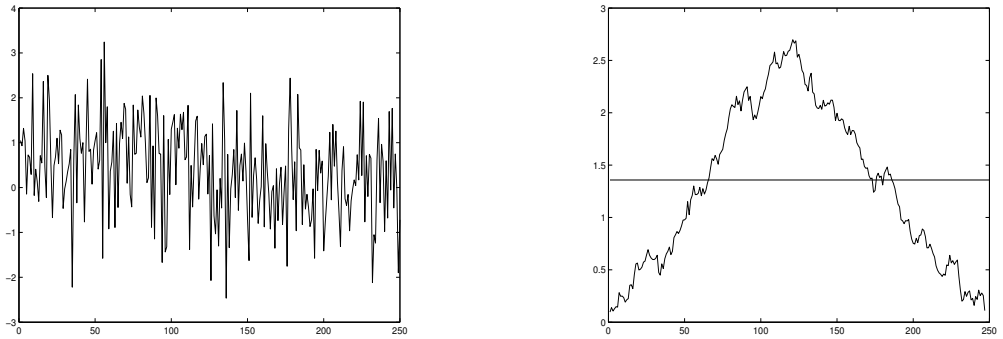
4.1 Behavior of the Statistic under Model Specification

For sake of illustration, we consider the following correctly specified model including a change

$$X_t = \begin{cases} 0.5 + (1 + \exp(0.5(1 + \beta_1 X_{t-1})))^{-1} + \varepsilon_t, & t \leqslant n/2, \\ 0.5 - (1 + \exp(0.5(1 + \beta_2 X_{t-1})))^{-1} + \varepsilon_t, & t > n/2, \end{cases}$$

where ε_t is i.i.d. standard Gaussian random variables.The estimation is carried out with a neural network for which $H = 1$.

As illustration, Figure 4.1 shows one generated data set in addition to the corresponding CUSUM-Plot showing $1/\sqrt{\widehat{\sigma}^2_{a,n}\,(n-p)}\,|\widehat{S}_n(k)|$Note that $T_n = \max_k 1/\sqrt{n-p}|\widehat{S}_n(k)|$.



(a) Computer based data (b) CUSUM-statistics

Figure 4.1: Neural Network based data and CUSUM statistic

Similar to the numerical study in the coming sections, we consider the sample sizes $n = 250, 500$ and base the empirical sizes and powers on 1000 replications (repeated

experiments). For $\beta_1 = 0.7$ and $\beta_2 = -0.7$ the size 0.048 (0.053, 0.042) for variance estimators (3.4) ((3.5) and the true variance) are obtained for a length of $n = 250$ (respectively 0.046, 0.053, 0.046 for $n = 500$) and corresponding power is 0.999, (0.796, 0.999) for $n = 250$ (respectively 0.999, 0.997, 0.999 for $n = 500$).

More generally consider

$$X_t = \begin{cases} 0.5 + (1 + \exp(0.5(1 + 0.7X_{t-1})))^{-1} + \varepsilon_t, & t \leqslant n/2, \\ \mu_2 + \alpha_2 (1 + \exp(0.5(1 + \beta_2 X_{t-1})))^{-1} + \varepsilon_t, & t > n/2. \end{cases}$$

For various values of $(\mu_2, \alpha_2, \beta_2)$ the empirical size and power of the test statistic are summarized in Table 4.1. The results here are given for a sample size of $n = 250$ observations, EV stands for variance estimator (3.4), LV for (3.5) and KV for the true variance of the errors.

			(0.1, 1, 0.7)	(0.5, -1, 0.7)	(0.5,1,-0.7)	(0.5, -1, -0.7)
		size	power	power	power	power
H=1	EV	0.048	0.790	0.999	0.160	1
	LV	0.053	0.697	0.7960	0.146	0.6810
	KV	0.042	0.786	0.999	0.146	1

Table 4.1: Influence of the of the various parameters on the empirical size for a nominal level of 5% for the correctly specified models, $n = 250$.

Conclusions:

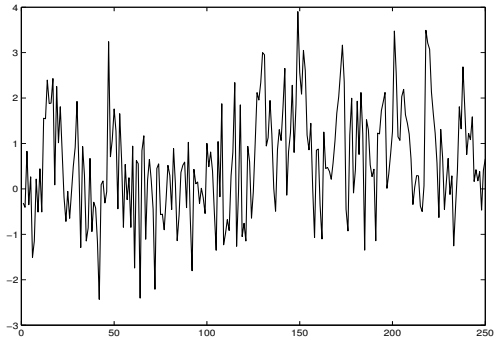
Using the estimator for the variance (3.4) yields comparable in fact even slightly better results than using the known variance which has been included as a benchmark value. Furthermore, the results are better especially in terms of power than using the long-run variance estimator. However, this is not surprising in the correctly specified model.

4.2 Behavior of the test statistics under misspecification

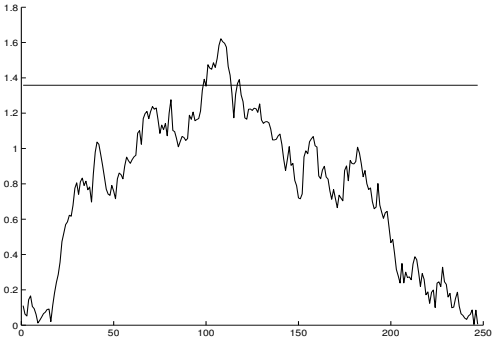
In this section we consider the more important misspecified situation, where the autoregression function is not truly given by a neural network but can only be approximated by it. In the simulations we consider linear autoregression functions $g(x) = a_0 + a_1 x$ corresponding to a true AR(1)-process, as well as nonlinear autoregression functions

$$g(x) = \begin{cases} a_0 + a_1 x, & x \leqslant c, \\ b_0 + b_1 x, & x > c, \end{cases} \quad \text{(TAR)}.$$

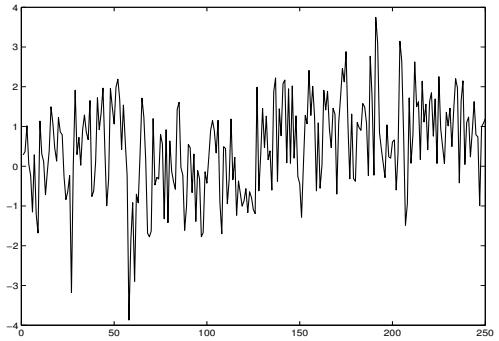
corresponding to a true TAR(1)-process (threshold autoregressive).



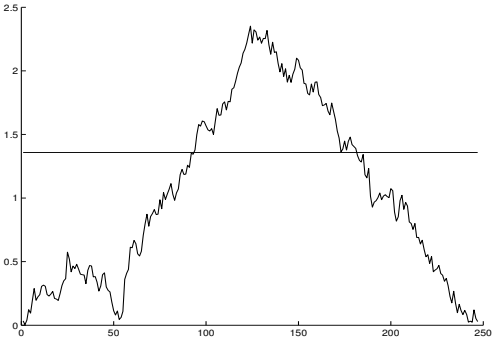
(a) AR 1: Sample Path



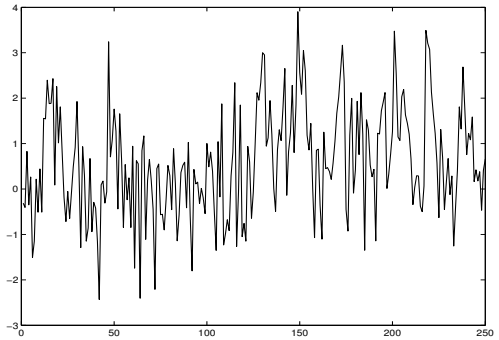
(b) AR 1: CUSUM Plot



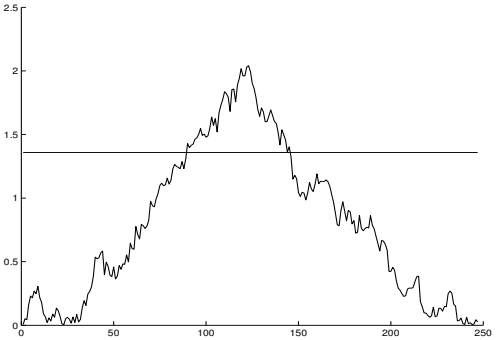
(c) AR 2: Sample Path



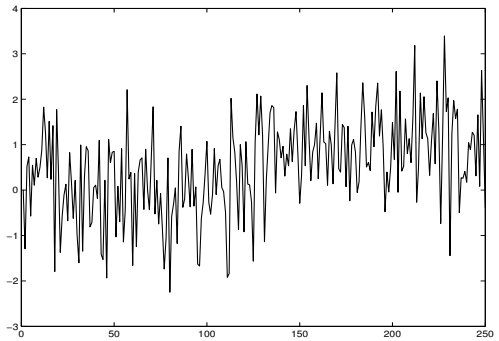
(d) AR 2: CUSUM Plot



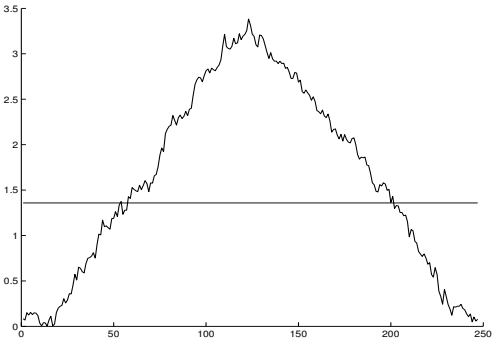
(e) TAR 1: Sample Path



(f) TAR 1: CUSUM Plot



(g) TAR 2: Sample Path



(h) TAR 2: CUSUM Plot

Figure 4.2: Sample Path and CUSUM plot for this sample path, $n = 250$

More precisely, we use standard normal errors and the following two AR as well as TAR parameters:

- AR 1: $g_0(x) = 0.3x$, $g_1(x) = 0.5 + 0.1x$
- AR 2: $g_0(x) = 0.3x$, $g_1(x) = 1 - 0.1x$
- TAR 1: $g_0(x) = 0.3x1_{\{x\geqslant 0\}} - 0.1x1_{\{x<0\}}$, $g_1(x) = (0.5 + 0.5x)1_{\{x\geqslant 0\}} - 0.3x1_{\{x<0\}}$
- TAR 2: $g_0(x) = 0.3x1_{\{x\geqslant 0\}} - 0.1x1_{\{x<0\}}$, $g_1(x) = (1-0.1x)1_{\{x\geqslant 0\}} + (0.5+0.1x)1_{\{x<0\}}$

From the construction of the test statistic it is clear that the variance estimator plays a crucial role for the performance of the test both under the null hypothesis as well as under alternatives (cf. Remark 3.1). The empirical size and power (based on 1000 repeated experiments) for different scenarios are reported in Table 4.2. EV indicates that the estimator $\widehat{\sigma}^2_{a,n}$ as given in 3.4 has been used, while for LV the long-run variance estimator (3.5) has been used. The true long-run variance is not known in the misspecified case so that it cannot be included here.

In Figure 4.2 one sample path for each four of the above misspecified scenarios is given in addition to a plot of $\frac{1}{\sqrt{\widehat{\sigma}^2_{a,n}\,(n-p)}}|\widehat{S}_n(k)|$ for this sample path, note that $T_n = \max_k 1/\sqrt{n-p}|\widehat{S}_n(k)|$.

			size	power
AR 1	$n = 250$	EV	0.035	0.819
		LV	0.041	0.751
	$n = 500$	EV	0.044	0.996
		LV	0.043	0.976
AR 2	$n = 250$	EV	0.035	1
		LV	0.041	0.934
	$n = 500$	EV	0.044	1
		LV	0.043	0.948
TAR 1	$n = 250$	EV	0.035	0.933
		LV	0.043	0.847
	$n = 500$	EV	0.041	0.999
		LV	0.043	0.981
TAR 2	$n = 250$	EV	0.037	0.956
		LV	0.039	0.837
	$n = 500$	EV	0.041	0.999
		LV	0.043	0.932

Table 4.2: Empirical size and power for a nominal 5% level of the test for several scenarios.

Conclusions:

The test is conservative in all cases and has a good power even in the misspecified situations. Using the variance estimator (3.4) instead of the long-run variance estimator (3.5)

results mainly in a smaller size but a larger power in our examples. These findings indicate that the errors of the approximating neural network autoregressive process are approximately independent so that the variance estimator (3.4) can be used and even yields superior results. Therefore, in the following the results are only given for this estimator.

4.3 Impact of Hidden Neurons:

Table 4.3 illustrates the influence of the number of hidden neurons on the power for the misspecified AR 1, AR 2, TAR 1 and TAR 2 models defined above. It can be seen that in all cases the power is good. For the AR 1 and AR 2 models the size is best for $H = 2$, while for the TAR 1 and TAR 2 models $H = 3$ delivers best results.

	AR 1		AR 2		TAR 1		TAR 2	
	size	power	size	power	size	power	size	power
H=2	0.046	0.989	0.046	1	0.040	0.998	0.040	1
H=3	0.042	0.994	0.042	1	0.041	0.999	0.041	1
H=4	0.034	0.991	0.034	1	0.028	1 0.999	0.028	1
H=5	0.033	0.994	0.033	1	0.035	1	0.035	1
H=6	0.032	0.993	0.032	1	0.041	0.999	0.041	0.998
H=10	0.036	0.987	0.036	1	0.024	0.998	0.024	1

Table 4.3: Influence of number of hidden neurons on empirical size and power for a nominal level of 5% for misspecified models, $n = 500$.

4.4 Power under Alternatives

Theorem 3.2 shows that certain alternatives are found with asymptotic power one. Remark 3.2 suggests that changes going along with a mean change will be more easily detectable, which is usually the case when statistics are based on estimated residuals. Therefore, we will have a closer look at this in simulations. Again, we consider the misspecified model, where the true process is generated by an AR(1)-process before the change-point and by a different AR(1)-process after the change-point.

Here, we consider the following four scenarios:

- AR 3: $g_0(x) = 1 + 0.5x$, $g_1(x) = 2$
- AR 4: $g_0(x) = 1 + \frac{2}{3}x$, $g_1(x) = 3$
- AR 5: $g_0(x) = 0.3x$, $g_1(x) = 0.9 - 0.8x$
- AR 6: $g_0(x) = 0.3x$, $g_1(x) = 1.5 - 0.5x$

For $H = 3$, the simulation results are summarized in Table 4.4.

	AR 3		AR 4		AR 5		AR 6	
	size	power	size	power	size	power	size	power
n=250	0.041	0.076	0.036	0.162	0.039	0.983	0.040	1
n=500	0.038	0.096	0.048	0.186	0.044	1.000	0.057	1

Table 4.4: Empirical size and power for a nominal level of 5% for misspecified models.

Note that $\mathbb{E}X_1^{(1)} = \frac{a_0}{1-a_1}$ and $\mathbb{E}X_1^{(2)} = \frac{b_0}{1-b_1}$, where $g_0(x) = a_0 + a_1 x$ and $g_1(x) = b_0 + b_1 x$. This shows that in for AR 3 and 4 it holds $\mathbb{E}X_1^{(1)} = \mathbb{E}X_1^{(2)}$, for AR 5 $|\mathbb{E}X_1^{(1)} - \mathbb{E}X_1^{(2)}| = 0.5$ and for AR 6 $|\mathbb{E}X_1^{(1)} - \mathbb{E}X_1^{(2)}| = 1$, where $\{X_t^{(1)}\}$ and $\{X_t^{(2)}\}$ denote the time series before and after the change respectively.

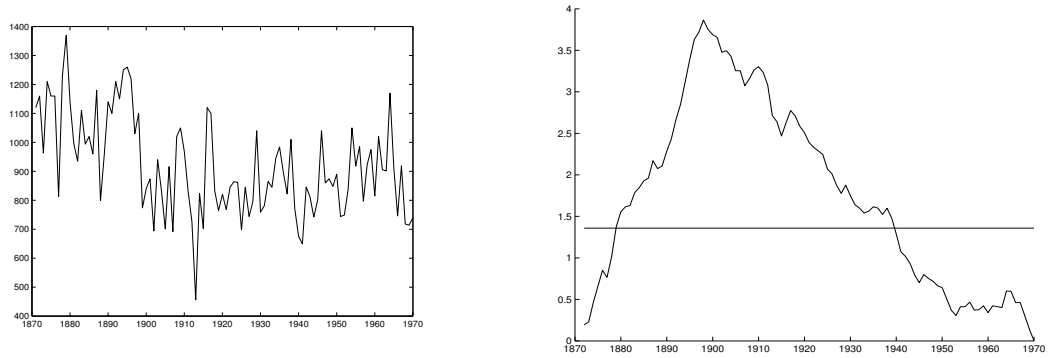
Conclusions:

As expected the power increases with increasing mean difference. However, in the cases with no mean difference at all AR 1 and 2, we still get an unbiased test with a power that is indeed significantly higher than the level (which is additionally given for comparison). Interestingly, AR 4 has almost twice the power of model AR 3.

4.5 Real Data Applications

In this section we apply our testing procedure to two real data sets that have been frequently used in change-point analysis (cf. e.g. Wu [55], Zhang et al. [56] or the R-package strucchange). The first one is a hydrological data set namely the Nile river flow recorded at Aswan (1871-1970), the second are the S&P-stock returns.

The Nile river data as well as a plot of $\frac{1}{\sqrt{\sigma_{a,n}^2 (n-p)}} |\widehat{S}_k|$ are given in Figure 4.3



(a) Raw data (b) CUSUM-statistics

Figure 4.3: Nile river data and CUSUM statistic

The CUSUM plot shows that the null hypothesis of no change is clearly rejected and the change-point is estimated to have occurred in 1898 (cf. Corollary 3.1). This is in

accordance to previous analyses of this data set. In fact, it was the year when the first Aswan dam was build.

An additional application of the test to the sub data sets before and after the estimated change-point did not yield any evidence of a second change-point.

The second data set we consider are the daily S&P log-returns from January 1992 to December 1999 (2022 observations) and July 1998 till June 2006 (2013 observations). The raw data sets, that consists of the daily index closing values, were downloaded from http://finance.yahoo.com/. S&P 500 is a world known stock index that is quoted at the New York stock exchange.

Instead of the log-returns r_t we apply the testing procedure to the log-transform of the squared returns $X_t = \log r_t^2$. A popular model for squared returns is given by the stochastic volatility model $r_t = \sigma_t Y_t$, where recently Stărică and Granger [47] have proposed that σ_t can be chosen piecewise constant while Y_t follow some weak autoregressive model (even i.i.d. noise is not such a bad assumption). The transformation $\log r_t^2 = \log \sigma_t^2 + \log Y_t^2$ brings the data rather close to our assumed model and transforms changes in the volatility into mean changes which can be detected quite well by our procedure as the simulation study has shown. However, in extreme cases where the values of the returns are close to zero, the logarithm of the squared returns are too small to fit reasonably into the framework. Therefore, we use a slight modification of the *log*-transformation as introduced in Fuller [19], page 496.

$$X_t^* = \log(r_t^2 + \iota\hat{\sigma}_r^2) - \frac{\iota\hat{\sigma}_r^2}{r_t^2 + \iota\hat{\sigma}_r^2} \tag{4.1}$$

where ι is a small real number (we choose $\iota = 0.02$) and $\hat{\sigma}_r^2$ is the sample variance of the returns. Figures 4.4 and 4.5 illustrate the raw data for the two data sets, the log-returns as well as their subsequent squared log-transformation.

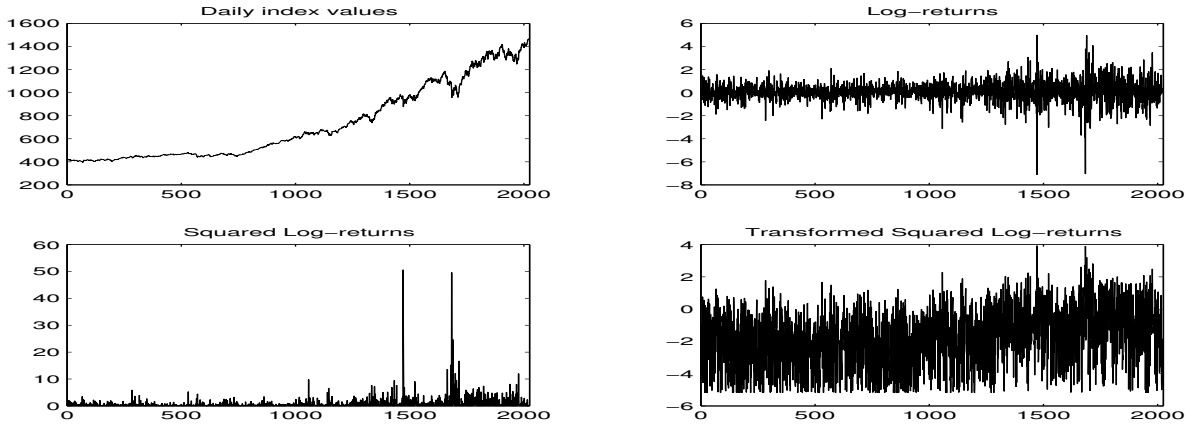


Figure 4.4: Daily S&P Values: January 1992- December 1999

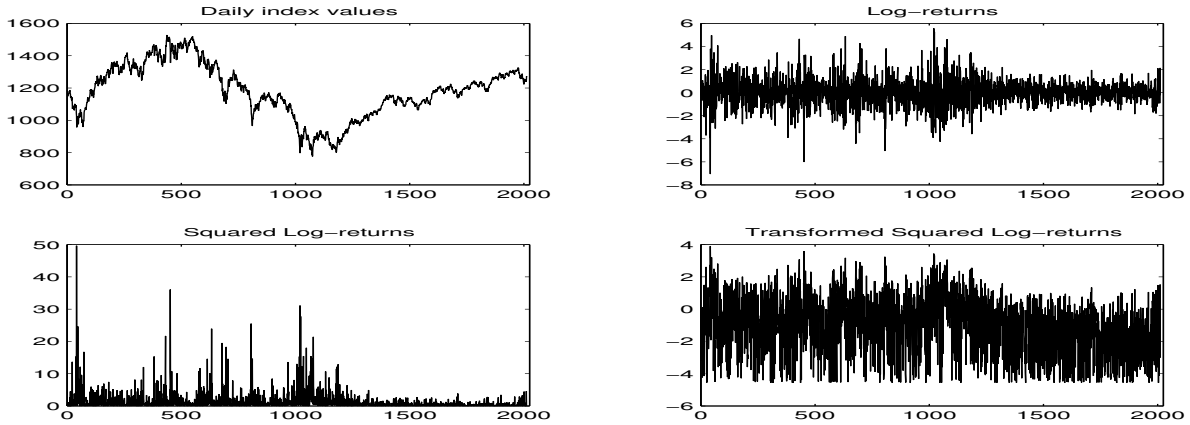
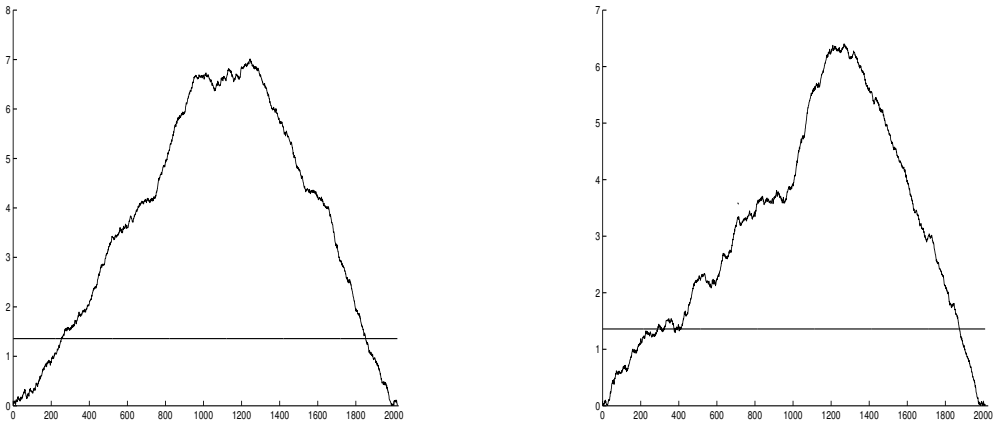


Figure 4.5: Daily S&P Values: July 1998- June 2006

The daily closing index values for both periods show the typical behavior of financial time series often referred to as 'stylized facts'. For example the returns exhibit some clustering and the squared returns consist of small positive values except for some local picks (large values). On the log-scale, the dynamic of the squared returns become more apparent.

Figure 4.6 shows the results of an application of our procedure to the log-transformed squared returns



(a) Jan 1992-Dec 1999 (b) July 1998-June 2006

Figure 4.6: CUSUM Plots for S&P

The CUSUM plots show that the null hypothesis of no change are clearly rejected and the change-points are estimated to have occurred around the time points 1246(5th December, 1996) for the first data and 1269 (23rd July, 2003) for the second data set. Both data sets have been investigated by Zhang et al [56] in the context of discriminating change points from long memory behavior. However, their procedures are based on the squared returns rather than the log-transformed squared returns nevertheless they find potential change-points close to our estimated change-points.

When applying change-point tests to financial data sets one has to be careful as it is well known that change-point tests reject in the presence of long-range dependence, while at the same time time series with structural breaks yield long-memory effects (cf. Andreou et al. [2] as well as Mikosch and Stărică [38].

5 Proofs of Section 2

We first state a uniform law of large numbers (ULLN) by Ranga Rao [46] for stationary and ergodic processes defined on a separable Banach space. As an application we obtain the uniform convergence of $Q_n(\theta)$ and its derivatives, which enables us to obtain the consistency of the NLLS estimator.

Theorem 5.1. *Let $|||$ be any norm on $\mathbb{R}^d$ and $v_t(\theta)$ be a stationary ergodic random sequence with values in $\mathbb{C}(K, \mathbb{R}^d)$ satisfying*

$$\mathbb{E} \sup_{\theta \in K} \|v_1(\theta)\| < \infty,$$

then

$$\sup_{\theta \in K} \left\| \frac{1}{n} \sum_{t=1}^{n} v_t(\theta) - \mathbb{E} v_1(\theta) \right\| \longrightarrow 0 \quad a.s. \ as \ n \longrightarrow \infty.$$

Proof. We refer to Theorem 6.5. in Ranga Rao [46]. ■

This enables us to prove the following uniform convergence theorem.

Proposition 5.1. *Let A.1 hold and K be compact.*

a) Then it holds as $n \to \infty$

$$\sup_{\theta \in K} \left| \frac{1}{n} Q_n(\theta) - E_\theta \right| \to 0 \quad a.s.,$$

where $Q_n(\theta) = \sum_{t=p+1}^{n} q_t(\theta)$ with $q_t(\theta)$ as in (1.5), E_θ as in (2.2).

b) As $n \to \infty$

$$\sup_{\theta \in K} \left\| \frac{1}{n} \nabla^2 Q_n(\theta) - \nabla^2 E_\theta \right\| \to 0 \quad a.s.,$$

where ∇^2 denotes the Hesse matrix with respect to θ.

A similar result can be formulated for the first partial derivative. However, we skip its presentation here as it is not needed for the proof of the main results in this paper.

Proof of Proposition 5.1. If X_t is stationary and ergodic (which is the case in the correctly specified as well as misspecified model without change), then

$$q_t(\theta) \qquad \text{and} \qquad \nabla^2 q_t(\theta)$$

are (pointwise for each θ) stationary and ergodic processes defined on $C(K, \mathbb{R})$ respectively $C(K, \mathbb{R}^{(H(p+2)+1)\times(H(p+2)+1)})$. Hence, preliminary conditions to make use of Theorem 5.1 are fulfilled. In case of a change at point $k^* = \lfloor \lambda n \rfloor$, this is still true under A.1 for $q_{t,1}(\theta) = (X_t^{(1)} - f(\mathbb{X}_{t-1}^{(1)}, \theta))^2$ as well as $q_{t,2}(\theta) = (X_t^{(2)} - f(\mathbb{X}_{t-1}^{(2)}, \theta))^2$ and an analogous expression for b). Noting that the mixed term is asymptotically negligible, it remains to show that $\mathbb{E}\sup_\theta q_1(\theta) < \infty$ and in case of a change that $\mathbb{E}\sup_\theta q_n(\theta) < \infty$ and corresponding expressions for b).

For a) this follows by $\sup_\theta |f(\mathbf{x}, \theta)| \leqslant D_1$ for some $D_1 > 0$ and Assumption A.1. An application of Theorem 5.1 to $\{q_t : t = 1, \ldots, \lfloor \lambda n \rfloor\}$ as well as $\{q_t : t = \lfloor \lambda n \rfloor + 1, \ldots, n\}$ yields the assertion in case of a change-point at $k^* = \lfloor \lambda n \rfloor$.

For b) it follows by the existence of second moments of the time series before and after the change because $\|\nabla^2 f(\mathbf{x}, \theta)\| \leqslant D_2 \max_{i=1,\ldots,p} x_i^2$. This also shows that by the dominated convergence theorem one can exchange taking derivatives and expectations, showing that the expectation is indeed given by $\nabla^2 E_\theta$. ■

The above proposition now enables us to prove Theorem 2.1.

Proof of Theorem 2.1. The assertion follows from Lemma 3.1 in Pötscher and Prucha [45] by Proposition 5.1 a) in addition to the identifiability uniqueness condition A.3. ■

Proof of Theorem 2.2. If X_t is α-mixing, then $(X_t, \mathbb{X}_{t-1})$ is α-mixing with the same rate as

$$\alpha_{(X_1, \mathbb{X}_0)}(k) \leqslant \alpha_X(k - p) \quad \text{for} \quad k - p > 0.$$

$\nabla q_t(\theta)$ is a measurable function of $(X_t^{(l)}, \mathbb{X}_{t-1}^{(l)})$, $l = 1, 2$, by definition of $f(\mathbf{x}, \theta)$, hence by an application of A.4 is also α-mixing with the same rate as $\{X_t^{(l)}\}$.

We first consider the case, where no change occurs: Since $\widetilde{\theta}_0 = \arg\min_\theta \mathbb{E} q_1(\theta)$ we get by A.3 that

$$0 = \nabla \mathbb{E} q_1(\widetilde{\theta}_0) = \mathbb{E} \nabla q_1(\widetilde{\theta}_0),$$

where we can exchange the limits by the dominated convergence theorem because $\sup_\theta |\nabla f(\mathbf{x}, \theta)| \leqslant D \max(|x_1|, \ldots, |x_p|)$ for some $D > 0$, hence by the mean value theorem a integrable majorant exists. By the strong invariance principle of Kuelbs and Philipp [33] for mixing sequences fulfilling A.4 the existence of V as well as the central limit theorem

$$\frac{1}{\sqrt{n}} \nabla Q_n(\widetilde{\theta}_0) \xrightarrow{\mathcal{D}} N(0, V) \tag{5.1}$$

holds - for a different proof of such a central limit theorem in the univariate case we refer to Oodaira and Yoshihara [41].

In case of a change assertion (5.1) follows by considering the central limit theorems before and after the change separately. By the independence of the process before and

after the change the joint central limit theorem is obtained on noting that the joint expectation is zero due to the definition of $\widetilde{\theta}_0$ and the joint variance is the sum of the variances due to the independence of the processes before and after the change.

By a Taylor expansion there exists θ_n^* such that $\|\theta_n^* - \widetilde{\theta}_0\| \leqslant \|\widehat{\theta}_n - \widetilde{\theta}_0\|$ with

$$\begin{aligned} 0 &= \nabla Q_n(\widehat{\theta}_n) \\ &= \nabla Q_n(\widetilde{\theta}_0) + (\widehat{\theta}_n - \widetilde{\theta}_0)\nabla^2 Q_n(\theta_n^*). \end{aligned}$$

Hence,

$$\nabla Q_n(\theta_0) = -(\widehat{\theta}_n - \widetilde{\theta}_0)\nabla^2 Q_n(\theta_n^*)$$

By Theorem 2.1 $\theta_n^* \overset{a.s.}{\to} \widetilde{\theta}_0$. Since $\widetilde{\theta}_0$ is an interior point of K we obtain that θ_n^* is $(a.s.)$ an interior point of K for n large enough. Thus Proposition 5.1 together with the Dominated Convergence Theorem implies as $\sup_\theta |\nabla^2 f(\mathbf{x}, \theta)| \leqslant D_3 \max(x_1^2, \ldots, x_p^2)$

$$\frac{1}{n}\nabla^2 Q_n(\theta_n^*) \to \nabla^2 E_{\widetilde{\theta}_0} \qquad a.s.,$$

which is positive definite by Assumption A.5. This yields the assertion by (5.1). ■

6 Proofs of Section 3

To prove Theorem 3.1 the following lemma plays the crucial role. It allows to replace the estimated errors by the centered errors, for which the assertions of Theorem 3.1 are well-known.

Lemma 6.1. *Let A.1 – A.5 hold and define* $\zeta_t = Z_t - f(\mathbb{Z}_{t-1}, \widetilde{\theta}_0)$. *Then it holds*

$$a) \qquad \max_{p<k\leqslant n} \sqrt{\frac{n-p}{k(n-p-k)}} \left| \sum_{t=p+1}^{k} (\widehat{\varepsilon}_t - (\zeta_t - \bar{\zeta}_{n-p})) \right| = O_P\left(\sqrt{\frac{\log\log n}{n}}\right),$$

$$b) \qquad \max_{p+G<k\leqslant n} \frac{1}{\sqrt{G}} \left| \sum_{t=k-G+1}^{k} (\widehat{\varepsilon}_t - (\zeta_t - \bar{\zeta}_{n-p})) \right| = O_P\left(\sqrt{\frac{\log\log n}{G}}\right),$$

where $\bar{\zeta}_n = 1/(n-p)\sum_{t=p+1}^{n} \zeta_t$. *In the correctly specified model without change* (1.4) *it holds* $\zeta_t = \varepsilon_t$.

Proof. By Theorem 2.1 it holds *a.s.* that $\widehat{\theta}_n \in K$ for n large enough, which implies (1.6). Hence,

$$\begin{aligned} &\sum_{t=p+1}^{k} (\widehat{\varepsilon}_t - (\zeta_t - \bar{\zeta}_n)) = \sum_{t=p+1}^{k} (\widehat{\varepsilon}_t - \zeta_t) - \frac{k}{n-p}\sum_{t=p+1}^{n} (\widehat{\varepsilon}_t - \zeta_t) \\ &= \sum_{t=p+1}^{k} (f(\mathbb{X}_{t-1}, \theta_0) - f(\mathbb{X}_{t-1}, \widehat{\theta}_n)) - \frac{k}{n-p}\sum_{t=p+1}^{n} (f(\mathbb{X}_{t-1}, \theta_0) - f(\mathbb{X}_{t-1}, \widehat{\theta}_n)) \qquad (6.1) \end{aligned}$$

A Taylor expansion of f yields

$$\begin{aligned} f(\mathbb{X}_{t-1},\widehat{\theta}_n) - f(\mathbb{X}_{t-1},\widetilde{\theta}_0) &= \nabla f(\mathbb{X}_{t-1},\widetilde{\theta}_0)^T(\widehat{\theta}_n - \widetilde{\theta}_0) \\ &\quad + \frac{1}{2}(\widehat{\theta}_n - \widetilde{\theta}_0)^T \nabla^2 f(\mathbb{X}_{t-1},\xi)(\widehat{\theta}_n - \widetilde{\theta}_0), \end{aligned} \tag{6.2}$$

where $\nabla f(\mathbb{X}_{t-1},\theta)$ is the gradient with respect to θ and $\nabla^2 f(\mathbb{X}_{t-1},\theta)$ is the Hessian matrix, $\widetilde{\theta}_0 < \xi < \widehat{\theta}_n$ elementwise. Furthermore the Hessian matrix is by the compactness of K uniformly bounded by $O(1)$ $\max_{1\leqslant i\leqslant p}\max_{1\leqslant j\leqslant p}|X_{t-i}X_{t-j}|$ similarly as in the proof of Theorem 2.2. The uniform ergodic theorem 5.1 yields

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

where $\|(\alpha_{i,j})\|_\infty = \max_{i,j}|\alpha_{i,j}|$. Together with (6.2) this yields uniformly in k

$$\begin{aligned} &\sum_{t=p+1}^{k}(f(\mathbb{X}_{t-1},\widehat{\theta}_n) - f(\mathbb{X}_{t-1},\widetilde{\theta}_0)) \\ &= \sum_{t=p+1}^{k}\nabla f(\mathbb{X}_{t-1},\widetilde{\theta}_0)^T(\widehat{\theta}_n - \widetilde{\theta}_0) + O_P\left(k\|\widehat{\theta}_n - \widetilde{\theta}_0\|^2\right). \end{aligned} \tag{6.3}$$

Assumption A.1 and the compactness of K show that $\mathbb{E}|\nabla f(\mathbb{X}_{t-1},\widetilde{\theta}_0)|^\nu < \infty$, since $\|\nabla f(\mathbb{X}_{t-1},\widetilde{\theta}_0)\|_\infty = O(\max_{1\leqslant j\leqslant k}|X_{t-j}|)$, where $\|(a_i)\|_\infty = \max_i|a_i|$. As in the proof of Theorem 2.2 Assumption A.4 shows that $\nabla f(\mathbb{X}_{t-1},\widetilde{\theta}_0)$ is mixing with exponential rate, hence by Kuelbs and Philipp [33] an invariance principle analogous to (3.3) holds, which implies the following law of iterated logarithm

$$\sum_{t=p+1}^{k}(\nabla f(\mathbb{X}_{t-1},\theta_0)^T - \mathbb{E}\nabla f(\mathbb{X}_{t-1},\theta_0)^T) = O(\sqrt{k\log\log k}) \qquad a.s. \tag{6.4}$$

A different proof of a law of iterated logarithm for mixing sequences but in the univariate situation is given by Oodaira and Yoshihara [40]. Together with Theorem 2.2 (6.1), (6.3) and (6.4) yield

$$\max_{p<k\leqslant n}\frac{1}{\sqrt{k}}\left|\sum_{t=p+1}^{k}(\widehat{\varepsilon}_t - (\zeta_t - \bar{\zeta}_n))\right| = O_P\left(\sqrt{\frac{\log\log n}{n}}\right).$$

Similar arguments yield the assertion if we replace $1/\sqrt{k}$ by $1/\sqrt{n-p-k}$ since by (6.1)

$$\sum_{t=p+1}^{k}(\widehat{\varepsilon}_t - (\zeta_t - \bar{\zeta}_n)) = \sum_{t=k+1}^{n}(\widehat{\varepsilon}_t - (\zeta_t - \bar{\zeta}_n)).$$

This proves assertion a). Similar arguments yield that

$$\sum_{t=k-G+1}^{k}(\widehat{\varepsilon}_t - (\zeta_t - \bar{\zeta}_{n-p})) = O_P\left(\sqrt{\log\log n}\right)$$

uniformly in k, thus assertion b) follows. ■

Proof of Theorem 3.1. Lemma 6.1 shows that we can replace $\widehat{\varepsilon}_i$ in the statistics by $\zeta_i - \bar{\zeta}_{n-p}$. Concerning T_{n2} note that

$$\frac{t(1-t)}{q(t)} = O(1),$$

by Csörgő and Horváth [11] (Chapter 4, Corollary 1.2). By the invariance principle (3.3) one can replace ζ_i by i.i.d. normal random variables with mean 0 and variance τ^2 by standard arguments from change-point analysis which are sketched below. The results then follow by classical results where $\widehat{\varepsilon}(t)$ in the statistics is replaced by $\xi(t) - \bar{\xi}_{n-p}$ and $\{\xi_i\}$ i.i.d. $N(0, \tau^2)$. For the proofs and further references we refer to the book by Csörgő and Horváth [12]. The result for d) can be found in Chen [9].

W.l.o.g. let $\mathbb{E}\zeta_1 = 0$. The invariance principle (3.3) implies a law of iterated logarithm from which we can deduce

$$\max_{1 \leqslant k \leqslant \log n} \sqrt{\frac{n}{k(n-k)}} \left| \sum_{t=p+1}^{k} (\zeta_t - \bar{\zeta}_{n-p}) \right| = o_P\left(\frac{\beta(\log n)}{\alpha(\log n)} \right)$$

and an analogous expression for $k \geqslant n - \log n$ as well as for $\sum_{t=p+1}^{k} (\xi_t - \bar{\xi}_{n-p})$. This shows that

$$\begin{aligned} &P\left(\alpha(\log n) \max_{1 \leqslant k \leqslant n} \sqrt{\frac{n}{k(n-k)}} \left| \sum_{t=p+1}^{k} (\zeta_t - \bar{\zeta}_{n-p}) \right| - \beta(\log n) \leqslant y \right) \\ &= P\left(\alpha(\log n) \max_{\log n \leqslant k \leqslant n - \log n} \sqrt{\frac{n}{k(n-k)}} \left| \sum_{t=p+1}^{k} (\zeta_t - \bar{\zeta}_{n-p}) \right| - \beta(\log n) \leqslant y \right) + o(1). \end{aligned}$$

Another application of the invariance principle shows that

$$\max_{\log n \leqslant k \leqslant n - \log n} \sqrt{\frac{n}{k(n-k)}} \left| \sum_{t=p+1}^{k} (\zeta_t - \bar{\zeta}_{n-p}) - \sum_{t=p+1}^{k} (\xi_t - \bar{\xi}_{n-p}) \right| = o_P(1)$$

finishing the proof for T_{n1}. Noting that by Csörgő and Horváth [11] (Chapter 4, Corollary 4) $\lim_{c_n \to 0} \sup_{t \leqslant c_n} \frac{\sqrt{t}}{q(t)} = 0$ the results for the other statistics follow similarly. ■

Proof of Lemma 3.1. First let us recall, from Theorem 2.2, that

$$\widehat{\theta}_n - \theta_0 = O_p\left(\frac{1}{\sqrt{n}} \right). \tag{6.5}$$

Then, we rewrite

$$\begin{aligned} \sum_{t=p+1}^{n} \widehat{\varepsilon}_t^2 &= \sum_{t=p+1}^{n} \varepsilon_t^2 - 2 \sum_{t=p+1}^{n} \varepsilon_t (f(\mathbb{X}_{t-1}, \widehat{\theta}_n) - f(\mathbb{X}_{t-1}, \theta_0)) \\ &\quad + \sum_{t=p+1}^{n} (f(\mathbb{X}_{t-1}, \widehat{\theta}_n) - f(\mathbb{X}_{t-1}, \theta_0))^2. \end{aligned} \tag{6.6}$$

From Marcinkiewicz-Zygmund (cf. e.g. Chow and Teicher [10], Theorem 5.2.2) we have for $2 \leqslant \nu < 4$

$$\frac{1}{n-(H(p+2)+1)} \sum_{t=p+1}^{n} \varepsilon_t^2 = \sigma^2 + o_P\left(n^{-(\nu-2)/\nu}\right)$$

respectively $O_P(n^{-1/2})$ by the central limit theorem if at least four moments exist. Analogously to (6.3) using only a first-order Taylor expansion we get by (6.5):

$$\begin{aligned} \frac{1}{n-(H(p+2)+1)} \sum_{t=p+1}^{n} (f(\mathbb{X}_{t-1}, \widehat{\theta}_n) - f(\mathbb{X}_{t-1}, \theta_0))^2 &= O_P\left(\left\|\widehat{\theta}_n - \theta_0\right\|^2\right) \\ &= O_P\left(\frac{1}{n}\right). \end{aligned}$$

Concerning the mixed term the Cauchy-Schwartz inequality yields

$$\frac{1}{n-(H(p+2)+1)} \sum_{t=p+1}^{n} \varepsilon_t (f(\mathbb{X}_{t-1}, \widehat{\theta}_n) - f(\mathbb{X}_{t-1}, \theta_0)) = O_P\left(n^{-1/2}\right),$$

which finishes the proof by (6.6). ■

Proof of Theorem 3.2. We can assume w.l.o.g. that $\widehat{\theta}_n$ is in the interior of K since otherwise we reject the null hypothesis right away. Hence, by (1.6) it holds

$$\hat{S}_n(k^*) = \sum_{j=p+1}^{k^*} (X_j - f(\mathbb{X}_{j-1}, \widehat{\theta}_n)) = -\sum_{j=k^*+1}^{n} (X_j - f(\mathbb{X}_{j-1}, \widehat{\theta}_n)). \tag{6.7}$$

This yields on the one hand

$$\begin{aligned} \hat{S}_n(k^*) &= \sum_{i=p+1}^{k^*} (Z_i - f(\mathbb{Z}_{i-1}, \widehat{\theta}_n)) \\ &= k^* \left(\mathbb{E}Z_1 - \mathbb{E}f(\mathbb{Z}_p, \theta)|_{\theta=\widehat{\theta}_n}\right) + \sum_{i=p+1}^{k^*} (Z_i - \mathbb{E}Z_1) \\ &\quad + O\left(\sup_{\theta \in K} \left| \sum_{i=p+1}^{k^*} (f(\mathbb{Z}_{i-1}, \theta) - \mathbb{E}f(\mathbb{Z}_p, \theta)) \right| \right) \\ &= \lambda n \left(\mathbb{E}Z_1 - \mathbb{E}f(\mathbb{Z}_p, \theta)|_{\theta=\widehat{\theta}_n}\right) + o_P(n) \end{aligned}$$

since by Theorem 5.1 a uniform law of large numbers holds (note that $\sup_x \sup_{\theta \in K} |f(\mathbf{x}, \theta)| = O(1)$).

On the other hand we obtain similarly (for $\mathbb{Y}_j = (Y_j, \ldots, Y_{j-p})$) using (6.7)

$$\hat{S}_n(k^*) = -(1-\lambda)n\left(\mathbb{E}Y_1 - \mathbb{E}f(\mathbb{Y}_p, \theta)|_{\theta=\widehat{\theta}_n}\right) + o_P(n),$$

since $f(\mathbb{Y}_j, \theta)$ is stationary and ergodic by Assumption $MS.1$. Together we obtain by Assumption A.6 (a)

$$\left|\hat{S}_n(k^*)\right| \geqslant nC\min(\lambda, 1-\lambda) + o_P(n) \tag{6.8}$$

for some constant $C > 0$, which implies immediately that

$$(\log\log n)^{-1/2}\, T_{n1} \xrightarrow{P} \infty,$$

which implies in return assertion a) (i). Furthermore, due to Assumption A.2 and since $q \in Q_{0,1}$ equation (6.8) implies

$$T_{n2}(q) \xrightarrow{P} \infty,$$

hence a) (ii). Concerning a) (iii) similar arguments yield

$$\begin{aligned}\left|\hat{S}_n(k^*) - \hat{S}_n(k^* - G)\right| &= G\left|\mathbb{E}Z_1 - \mathbb{E}f(\mathbb{Z}_p, \theta)|_{\theta=\widehat{\theta}_n}\right| + o_P(G)\\ \left|\hat{S}_n(k^* + G) - \hat{S}_n(k^*)\right| &= G\left|\mathbb{E}Y_1 - \mathbb{E}f(\mathbb{Y}_p, \theta)|_{\theta=\widehat{\theta}_n}\right| + o_P(G),\end{aligned}$$

hence by Assumption A.6 (a)

$$T_{n3}(G) \geqslant \sqrt{G}C + o_P(\sqrt{G})$$

for some $C > 0$, which in turn yields $(\log(n/G))^{-1/2} T_{n3}(G) \xrightarrow{P} \infty$ and hence the assertion. Concerning a) (iv) we need again a slight variation of the argument, namely it holds uniformly in k

$$\left|\hat{S}_n(k)\right| = \begin{cases} k\left|\mathbb{E}Z_1 - \mathbb{E}f(\mathbb{Z}_p, \theta)|_{\theta=\widehat{\theta}_n}\right| + o_P(k), & n/\log n < k \leqslant k^*,\\ (n-k)\left|\mathbb{E}Y_1 - \mathbb{E}f(\mathbb{Y}_p, \theta)|_{\theta=\widehat{\theta}_n}\right| + o_P(n-k), & k^* < k < n - n/\log n. \end{cases}$$

Since by (3.1)

$$\frac{1}{n}\sum_{k^* \geqslant k \geqslant \frac{n}{\log n}} \frac{k^2/n}{r(k/(n-p))} \geqslant \frac{n}{\log n}\left(\int_0^\lambda \frac{t}{r(t)}\, dt + o(1)\right)$$

and an analogous expression for $k^* \leqslant j \leqslant n - n/\log n$, we obtain by Assumption A.6 (a)

$$T_{n4}(r) \geqslant o_P\left(\frac{n}{\log n}\right) + C\frac{n}{\log n}$$

for some $C > 0$ and hence the assertion.

Similarly to $T_{n3}(G)$ we obtain for b) that

$$\begin{aligned}\widetilde{T}_{n3}(G) &\geqslant \frac{1}{\sqrt{2G}}\left|\hat{S}_n(k^*+G)-2\hat{S}_n(k^*)+\hat{S}_n(k^*-G)\right| \\ &\geqslant \sqrt{G/2}\left|(\mathbb{E}Z_1-\mathbb{E}f(\mathbb{Z}_p,\theta)|_{\theta=\widehat{\theta}_n})-(\mathbb{E}Y_1-\mathbb{E}f(\mathbb{Y}_p,\theta)|_{\theta=\widehat{\theta}_n})\right|+o_P(\sqrt{G}),\end{aligned}$$

and hence the assertion. ■

Proof of Corollary 3.1. The proof of Theorem 3.2 yields that

$$\sup_{t\in[0,1]}\left|\frac{1}{n}\left|\hat{S}_n(\lfloor nt\rfloor)\right|-L_n(t)\right|=o_P(1),$$

where

$$L_n(t)=\begin{cases}t\left|\mathbb{E}Z_1-\mathbb{E}f(\mathbb{Z}_p,\theta)|_{\theta=\widehat{\theta}_n}\right|, & t<\lambda,\\ \max\left(\lambda\left|\mathbb{E}Z_1-\mathbb{E}f(\mathbb{Z}_p,\theta)|_{\theta=\widehat{\theta}_n}\right|,(1-\lambda)\left|\mathbb{E}Y_1-\mathbb{E}f(\mathbb{Y}_p,\theta)|_{\theta=\widehat{\theta}_n}\right|\right), & t=\lambda,\\ (1-t)\left|\mathbb{E}Y_1-\mathbb{E}f(\mathbb{Y}_p,\theta)|_{\theta=\widehat{\theta}_n}\right|, & t>\lambda.\end{cases}$$

$L_n(t)$ has a unique maximum in $t=\lambda$ for all n, is equicontinuous (with respect to n) for all $t\neq\lambda$ (since $f(x,\theta)$ is bounded). Let $\xi_n:=\max(|\mathbb{E}f(\mathbb{Z}_p,\theta)|_{\theta=\widehat{\theta}_n}-\mathbb{E}Z_1|,|\mathbb{E}Y_1-\mathbb{E}f(\mathbb{Y}_p,\theta)|_{\theta=\widehat{\theta}_n}|)$ with $P(\xi_n\geqslant c)\to 1$ by Assumption A.6 (a). Since $\xi\geqslant c-(c-\xi)_+$ and $P\left((c-\xi)_+\geqslant\varepsilon\right)\leqslant P(\xi<c)=1-P(\xi\geqslant c)\to 0$, it holds $\xi_n\geqslant c+o_P(1)$. We can now show that

$$\inf_n(L_n(\lambda)-L_n(t))>c(t,\lambda)+o_P(1),\qquad c(t,\lambda)>0\quad\text{for any } t\neq\lambda$$

We prove this assertion for $t<\lambda$, the assertion for $t>\lambda$ is analogous. First, consider the case when $\widehat{\theta}_n$ is such that $|\mathbb{E}f(\mathbb{Z}_p,\theta)|_{\theta=\widehat{\theta}_n}-\mathbb{E}Z_1|\geqslant\min\left(\frac{1-\lambda}{\lambda},1\right)\xi_n$. In this case it can easily be seen that

$$L_n(\lambda)-L_n(t)\geqslant(\lambda-t)\,\xi_n\,\min\left(\frac{1-\lambda}{\lambda},1\right)\geqslant c(\lambda-t)\,\min\left(\frac{1-\lambda}{\lambda},1\right)+o_P(1).$$

Otherwise it holds $|\mathbb{E}f(\mathbb{Z}_p,\theta)-\mathbb{E}Z_1|<\min\left(\frac{1-\lambda}{\lambda},1\right)c$, hence by by definition of ξ_n that $|\mathbb{E}Y_1-\mathbb{E}f(\mathbb{Y}_p,\theta)|_{\theta=\widehat{\theta}_n}|\geqslant c$, which yields

$$L_n(\lambda)-L_n(t)\geqslant\xi_n\left[1-\lambda-t\min\left(\frac{1-\lambda}{\lambda},1\right)\right]\geqslant(\lambda-t)\,c\,\min\left(\frac{1-\lambda}{\lambda},1\right)+o_P(1),$$

proving the assertion.

From this we can conclude the proof. We only give the argument for real sequences $s_n(t)$ and $l_n(t)$ – the assertion for random sequences $\frac{1}{n}\left|\widetilde{S}_n(\lfloor nt\rfloor)\right|$ and $L_n(t)$ then follows via the subsequence principle. This variation of the standard proof is necessary since we do not know anything about the limit of $L_n(t)$ (not even if it exists): Suppose

$\widehat{\lambda}_n = \arg\max(s_n(t)) \not\to \lambda$. Because $[0,1]$ is compact there exists a subsequence $\widehat{\lambda}_{\alpha(n)}$ and $t_1 \neq \lambda$ with $\widehat{\lambda}_{\alpha(n)} \to t_1$, hence

$$|s_{\alpha(n)}(\widehat{\lambda}_{\alpha(n)}) - y_{\alpha(n)}(t_1)| \leqslant \max_t |s_{\alpha(n)}(t) - y_{\alpha(n)}(t)| + |y_{\alpha(n)}(\widehat{\lambda}_{\alpha(n)}) - y_{\alpha(n)}(t_1)| \to 0,$$

since by assumption $\sup_t |s_{\alpha(n)}(t) - l_n(t)| \to 0$ and by the equicontinuity of $l_n, n \in \mathbb{N}$. Since $\inf_n(l_n(\lambda) - l_n(t_1)) > c(t_1, \lambda) > 0$ we conclude

$$y_{\alpha(n)}(\lambda) - s_{\alpha(n)}(\widehat{\lambda}_{\alpha(n)}) = y_{\alpha(n)}(\lambda) - y_{\alpha(n)}(t_1) + y_{\alpha(n)}(t_1) - s_{\alpha(n)}(\widehat{\lambda}_{\alpha(n)}) > c(t_1, \lambda) + o(1),$$

but this is a contradiction since by assumption it holds

$$|s_n(\widehat{\lambda}_n) - l_n(\lambda)| = |\sup_t s_n(t) - \sup_t l_n(t)| \leqslant \sup_t |s_n(t) - l_n(t)| \to 0.$$

■

References

[1] An, H. Z. and Huang, F. C. The geometrical ergodicity of nonlinear autoregressive models. *Statist. Sinica*, 6:943–956, 1996.

[2] Andreou, E., and Ghysels, E. Structural breaks in financial time series. Andersen, Torben G. (ed.) et al., Handbook of financial time series. With a foreword by Robert Engle. Berlin: Springer. 839-870 (2009)., 2009.

[3] Andrews, D. W. K. Test for parametric instability and structural change with unknown change point. *J. Theoret. Probab.*, 61:821–856, 1993.

[4] J. Bai. Least squares estimation of a shift in linear processes. *J. Time Ser. Anal.*, 15:435–472, 1994.

[5] Berkes, I. and Horváth, L. Limit results for the empirical process of squared residuals in GARCH models. *Stoch. Process. Appl.*, 105:271–298, 2003.

[6] Berkes, I., Horváth, L., and Kokoszka, P. Asymptotics for GARCH squared residual correlations. *Econom. Theory*, 19:515–540, 2003.

[7] Berkes, I., Horváth, L., and Kokoszka, P. Testing for parameter constancy in GARCH(p, q) models. *Stat. Probab. Lett.*, 70(4):263–273, 2005.

[8] Brown, R. L., Durbin, J., and Evans, J. M. Techniques for testing the constancy of regression relationships over time. *J. R. Stat. Soc. Ser. B*, 37:149–163, 1975.

[9] Chen, X. Inference in a simple change-point problem. *Scientia Sinica A*, 31:654–667, 1998.

[10] Chow, Y.S., and Teicher, H. *Probability Theory – Independence, Interchangeability, Martingales*. Springer, New York, third edition, 1997.

[11] Csörgő, M., and Horváth, L. *Weighted Approximations in Probability and Statistics*. Wiley, Chichester, 1993.

[12] Csörgő, M., and Horváth, L. *Limit Theorems in Change-Point Analysis.* Wiley, Chichester, 1997.

[13] Davis, R.A., Huang, D., and Yao, Y.-C. Testing for a change in the parameter values and order of an autoregressive model. *Ann. Statist.*, 23:282–304, 1995.

[14] Delgado, M. A. and Hidalgo, J. Nonparametric inference on structural breaks. *J. Econometrics*, 96:113–144, 2000.

[15] Doukhan, P. and S. Louhichi. A new weak dependence condition and applications to moment inequalities. *Stochastic Processes and their Applications*, 84:313–342, 1999.

[16] Engle, R. Autoregressive conditional hetereskedasticity with estimates of United Kingdom inflation. *Econometrica*, 50:987–1008, 1982.

[17] Francq, C., Horváth, L., and Zakoian, J. Sup-tests for linearity in a general nonlinear ar(1) model. *Econom. Theory*, 26:965–993, 2010.

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

[19] Fuller, W. A. *Introduction to statistical time series.* Wiley, New York, second edition, 1996.

[20] Guégan, D. and Diebolt, J. Probabilistic properties of the β-ARCH model. *Statist. Sinica*, 4:71–87, 1994.

[21] Horváth, L. Change in autoregressive processes. *Stoch. Process. Appl.*, 44:221–242, 1993.

[22] Horváth, L., Kokoszka, P., and Teyssiere, G. Empirical process of the squared residuals of an arch sequence. *Ann. Statist.*, 2:445–469, 2001.

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

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

[25] Hušková, M., Prášková, Z., and Steinebach, J. On the detection of changes in autoregressive time series, I. Asymptotics. *J. Statist. Plann. Infer.*, 137:1243–1259, 2007.

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

[27] Jensen, S. T. and Rahbek, A. On the law of large numbers for (geometrically) ergodic Markov chains. *Econom. Theory*, 23:761–766, 2007.

[28] Kirch, C. Resampling in the frequency domain of time series to determine critical values for change-point tests. *Statist. Decisions*, 25:237–261, 2007.

[29] Kirch, C. and Tadjuidje K., J. . An online approach to detecting changes in nonlinear autoregressive model. *Preprint*, 2011.

[30] Kirch, C. and Tadjuidje K., J. . A uniform central limit theorem for neural network based autoregressive processes with applications to change-point analysis. *Preprint*, 2011.

[31] 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.

[32] 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.

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

[34] Kulperger, R.J. On the residuals of autoregressive processes and polynomial regression. *Stoch. Process. Appl.*, 21:107–118, 1985.

[35] Luukkonen, P., Saikkonen, P., and Teräsvirta, T. Testing linearity against smooth transition autoregressive models. *Biometrika*, 75:491–499, 1988.

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

[37] Meyn, S.P. and Tweedie, R.L. . *Markov Chains and Stochastic Stability*. Springer, London, 1993.

[38] Mikosch, T. and Stărică, C. Nonstationarities in financial time series, the long-range dependence, and the igarch effects. *Rev. Econom. Statist.*, 86:378–390, 2004.

[39] Müller, H.-G. Change points in nonparametric regression analysis. *Ann. Statist.*, 20:737–761, 1992.

[40] Oodaira, H. and Yoshihara, K-I. The law of the iterated logarithm for stationary processes satisfying mixing conditions. *Kodai Math. Sem. Rep.*, 23:311–334, 1971.

[41] Oodaira, H. and Yoshihara, K-I. Functional central limit theorems for strictly stationary process satisfying the strong mixing condition. *Kodai Math. Sem. Rep.*, 24:259–269, 1972.

[42] Page, E.S. Continuous inspection schemes. *Biometrika*, 41:100–115, 1954.

[43] Page, E.S. Control charts with warning lines. *Biometrika*, 42:243–257, 1955.

[44] Politis, D. N. Adaptive bandwidth choice. *J. Nonparametr. Stat.*, 15:517–533, 2003.

[45] Pötscher, B. M. and Prucha, I. R. *Dynamic nonlinear econometric models. Asymptotic theory*. Springer, Berlin, 1997.

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

[47] Starica, C., and Granger, C. Nonstationarities in stock returns. *The Review of Economics and Statistics*, 87, 2005.

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

[49] Taniguchi, M. and Kakizawa, Y. *Asymptotic theory of statistical inference for time series*. Springer, New York, 2000.

[50] Teräsvirta, T., van Dijk, D., and Medeiros, M. Linear models, smooth transition autoregressions, and neural networks for forecasting macroeconomic time series: A reexamination. *International Journal of Forecasting*, 21:755–774, 2005.

[51] Tong, H. *Non-linear time series: a dynamical system approach.* Oxford University Press, 1990.

[52] White, H. Economic prediction using neural networks: The case of IBM daily stoch returns. *Proceedings of the 2nd annual IEEE conference on neural networks*, II:451–458, 1988.

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

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

[55] Wu, Y. *Inference for change point and post change mean after a CUSUM test.* Springer, New York, 2004.

[56] Zhang, A., Gabrys, R., and Kokoszka, P. Discriminating between long memory and volatility shifts. *Austrian Journal of Statistics*, 36:253–275, 2007.