Multiscale Inference and Long-Run Variance Estimation in Nonparametric Regression with Time Series Errors

Marina Khismatullina[1] University of Bonn

Michael Vogt[2] University of Bonn

September 14, 2018

In this paper, we develop multiscale methods to test qualitative hypotheses about the function m in the nonparametric regression model $Y_{t,T} = m(t/T) + \varepsilon_t$ with time series errors ε_t. In time series applications, m represents a nonparametric time trend. Practitioners are often interested in whether the trend function m has certain shape properties. For example, they would like to know whether m is constant or whether it is increasing/decreasing in certain time regions. Our multiscale methods allow to test for such shape properties of the trend m. In order to perform the tests, we require an estimator of the long-run error variance $\sigma^2 = \sum_{\ell=-\infty}^{\infty} \text{Cov}(\varepsilon_0, \varepsilon_\ell)$. We propose a new estimator of σ^2 for the case that $\{\varepsilon_t\}$ is a general AR(p) process. The main advantage of this estimator is that it is completely free of tuning parameters and very simple to compute. In the technical part of the paper, we derive asymptotic theory for the proposed multiscale tests and the estimator of the long-run error variance. The theory is complemented by a simulation study and two empirical examples from climatology and economics.

Key words: Multiscale statistics; long-run variance; nonparametric regression; time series errors; shape constraints; strong approximations; anti-concentration bounds.
AMS 2010 subject classifications: 62E20; 62G10; 62G20; 62M10.

1 Introduction

The analysis of time trends is an important aspect of many time series applications. In a wide range of situations, practitioners are particularly interested in certain shape properties of the trend. They raise questions such as the following: Does the observed time series have a trend at all? If so, is the trend increasing/decreasing in certain time regions? Can one identify the regions of increase/decrease? As an example, consider the time series plotted in Figure 1 which shows the yearly mean temperature in Central

[1]Address: Bonn Graduate School of Economics, University of Bonn, 53113 Bonn, Germany. Email: marina.k@uni-bonn.de.
[2]Corresponding author. Address: Department of Economics and Hausdorff Center for Mathematics, University of Bonn, 53113 Bonn, Germany. Email: michael.vogt@uni-bonn.de.

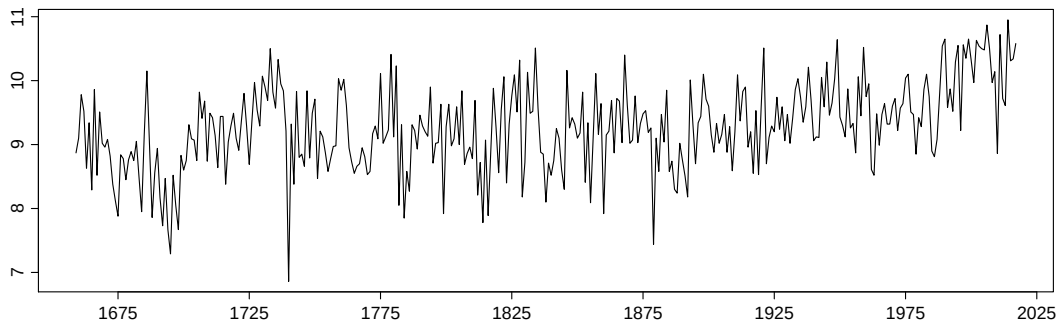

Figure 1: Yearly mean temperature in Central England from 1659 to 2017 measured in °C.

England from 1659 to 2017. Climatologists are very much interested in learning about the trending behaviour of temperature time series like this; see e.g. Benner (1999) and Rahmstorf et al. (2017). Among other things, they would like to know whether there is an upward trend in the Central England mean temperature towards the end of the sample as visual inspection might suggest.

In this paper, we develop new methods to test for certain shape properties of a nonparametric time trend. We in particular construct a multiscale test which allows to identify local increases/decreases of the trend function. We develop our test in the context of the following model setting: We observe a time series $\{Y_{t,T} : 1 \leq t \leq T\}$ of the form

$$Y_{t,T} = m\Big(\frac{t}{T}\Big) + \varepsilon_t \tag{1.1}$$

for $1 \leq t \leq T$, where $m : [0,1] \to \mathbb{R}$ is an unknown nonparametric regression function and the error terms ε_t form a stationary time series process with $\mathbb{E}[\varepsilon_t] = 0$ for all t. In a time series context, the design points t/T represent the time points of observation and m is a nonparametric time trend. As usual in nonparametric regression, we let the function m depend on rescaled time t/T rather than on real time t. A detailed description of model (1.1) is provided in Section 2.

Our multiscale test is developed step by step in Section 3. Roughly speaking, the procedure can be outlined as follows: Let $H_0(u,h)$ be the hypothesis that m is constant in the time window $[u-h, u+h] \subseteq [0,1]$, where u is the midpoint and $2h$ the size of the window. In a first step, we set up a test statistic $\widehat{s}_T(u,h)$ for the hypothesis $H_0(u,h)$. In a second step, we aggregate the statistics $\widehat{s}_T(u,h)$ for a large number of different time windows $[u-h, u+h]$. We thereby construct a multiscale statistic which allows to test the hypothesis $H_0(u,h)$ simultaneously for many time windows $[u-h, u+h]$. In the technical part of the paper, we derive the theoretical properties of the resulting multiscale test. To do so, we come up with a proof strategy which combines strong approximation results for dependent processes with anti-concentration bounds for Gaussian random vectors. This strategy is of interest in itself and may be applied to other multiscale test problems for dependent data. As shown by our theoretical

analysis, our multiscale test allows to make simultaneous confidence statements of the following form: For a given confidence level $\alpha \in (0,1)$, there is an increase/decrease in the trend m on all time windows $[u-h, u+h]$ for which the hypothesis $H_0(u,h)$ is rejected. Hence, the multiscale test allows to identify, with a pre-specified statistical confidence, time regions where the trend m is increasing/decreasing.

For independent data, multiscale tests have been developed in a variety of different contexts in recent years. In the regression context, Chaudhuri and Marron (1999, 2000) introduced the so-called SiZer method which has been extended in various directions; see e.g. Hannig and Marron (2006) where a refined distribution theory for SiZer is derived. Hall and Heckman (2000) constructed a multiscale test on monotonicity of a regression function. Dümbgen and Spokoiny (2001) developed a multiscale approach which works with additively corrected supremum statistics and derived theoretical results in the context of a continuous Gaussian white noise model. Rank-based multiscale tests for nonparametric regression were proposed in Dümbgen (2002) and Rohde (2008). More recently, Proksch et al. (2018) have constructed multiscale tests for inverse regression models. In the context of density estimation, multiscale tests have been investigated in Dümbgen and Walther (2008), Rufibach and Walther (2010), Schmidt-Hieber et al. (2013) and Eckle et al. (2017) among others.

Whereas a large number of multiscale tests for independent data have been developed in recent years, multiscale tests for dependent data are much rarer. Most notably, there are some extensions of the SiZer approach to a time series context. Park et al. (2004) and Rondonotti et al. (2007) have introduced dependent SiZer methods which can be regarded as an alternative to our multiscale test. However, whereas their analysis is mainly methodological, we back up our multiscale test by a complete asymptotic theory. Our multiscale approach is also related to Wavelet-based methods: It investigates the data on different intervals $[u-h, u+h]$. Similar to Wavelet-based methods, it thus takes into account different locations u and resolution levels h simultaneously. Nevertheless, we are not aware of any Wavelet-based test for local increases/decreases of the nonparametric trend function in model (1.1). Wavelet-based methods have been used for other purposes in the literature such as estimating/reconstructing nonparametric regression functions [see e.g. Donoho et al. (1995) or Von Sachs and MacGibbon (2000)] and change point detection [see e.g. Cho and Fryzlewicz (2012)].

Our multiscale test depends on the long-run error variance $\sigma^2 = \sum_{\ell=-\infty}^{\infty} \mathrm{Cov}(\varepsilon_0, \varepsilon_\ell)$, which is usually unknown in practice. To carry out the test, we thus require an estimator of σ^2. Indeed, such an estimator is required for virtually all inferential procedures in the context of model (1.1). Hence, the problem of estimating σ^2 in model (1.1) is of broader interest and has received a lot of attention in the literature; see Müller and Stadtmüller (1988), Herrmann et al. (1992) and Hall and Van Keilegom (2003) among many others. In Section 4, we discuss several estimators of σ^2 which are valid under different conditions on the error process $\{\varepsilon_t\}$. Most notably, we introduce a new

estimator of σ^2 for the case that $\{\varepsilon_t\}$ is an AR(p) process. The main advantage of this estimator is that it is completely free of tuning parameters and also very simple to compute.

The methodological and theoretical analysis of the paper is complemented by a simulation study in Section **??** and two empirical applications in Section **??**. In the simulation study, we examine the finite sample properties of our multiscale test and compare it to the dependent SiZer methods of Park et al. (2004) and Rondonotti et al. (2007). Moreover, we investigate the small sample performance of our estimator of σ^2 in the AR(p) case and compare it to the estimator of Hall and Van Keilegom (2003). In Section **??**, we apply our methods to the temperature data from Figure 1 and to a historic time series of economic inequality data.

2 The model

We now describe the model setting in detail which was briefly outlined in the Introduction. We observe a time series $\{Y_{t,T} : 1 \leq t \leq T\}$ of length T which satisfies the nonparametric regression equation

$$Y_{t,T} = m\Big(\frac{t}{T}\Big) + \varepsilon_t \tag{2.1}$$

for $1 \leq t \leq T$. Here, m is an unknown nonparametric function defined on $[0,1]$ and $\{\varepsilon_t : 1 \leq t \leq T\}$ is a zero-mean stationary error process. For simplicity, we restrict attention to equidistant design points $x_t = t/T$. However, our methods and theory can also be carried over to non-equidistant designs. The stationary error process $\{\varepsilon_t\}$ is assumed to have the following properties:

(C1) The variables ε_t allow for the representation $\varepsilon_t = G(\ldots, \eta_{t-1}, \eta_t, \eta_{t+1}, \ldots)$, where η_t are i.i.d. random variables and G is a measurable function.

(C2) It holds that $\|\varepsilon_t\|_q < \infty$ for some $q > 4$, where $\|\varepsilon_t\|_q = (\mathbb{E}|\varepsilon_t|^q)^{1/q}$.

Following Wu (2005), we impose conditions on the dependence structure of the error process $\{\varepsilon_t\}$ in terms of the physical dependence measure $d_{t,q} = \|\varepsilon_t - \varepsilon_t'\|_q$, where $\varepsilon_t' = G(\ldots, \eta_{-1}, \eta_0', \eta_1, \ldots, \eta_{t-1}, \eta_t, \eta_{t+1}, \ldots)$ with $\{\eta_t'\}$ being an i.i.d. copy of $\{\eta_t\}$. In particular, we assume the following:

(C3) Define $\Theta_{t,q} = \sum_{|s|\geq t} d_{s,q}$ for $t \geq 0$. It holds that

$$\Theta_{t,q} = O\big(t^{-\tau_q}(\log t)^{-A}\big),$$

where $A > \frac{2}{3}(1/q + 1 + \tau_q)$ and $\tau_q = \{q^2 - 4 + (q-2)\sqrt{q^2 + 20q + 4}\}/8q$.

The conditions (C1)–(C3) are fulfilled by a wide range of stationary processes $\{\varepsilon_t\}$. As a first example, consider linear processes of the form $\varepsilon_t = \sum_{i=0}^{\infty} c_i \eta_{t-i}$ with $\|\varepsilon_t\|_q < \infty$, where c_i are absolutely summable coefficients and η_t are i.i.d. innovations with $\mathbb{E}[\eta_t] = 0$ and $\|\eta_t\|_q < \infty$. Trivially, (C1) and (C2) are fulfilled in this case. Moreover, if $|c_i| = O(\rho^i)$ for some $\rho \in (0,1)$, then (C3) is easily seen to be satisfied as well. As a special case, consider an ARMA process $\{\varepsilon_t\}$ of the form $\varepsilon_t + \sum_{i=1}^{p} a_i \varepsilon_{t-i} = \eta_t + \sum_{j=1}^{r} b_j \eta_{t-j}$ with $\|\varepsilon_t\|_q < \infty$, where $a_1, \ldots, a_p$ and $b_1, \ldots, b_r$ are real-valued parameters. As before, we let η_t be i.i.d. innovations with $\mathbb{E}[\eta_t] = 0$ and $\|\eta_t\|_q < \infty$. Moreover, as usual, we suppose that the complex polynomials $A(z) = 1 + \sum_{j=1}^{p} a_j z^j$ and $B(z) = 1 + \sum_{j=1}^{r} b_j z^j$ do not have any roots in common. If $A(z)$ does not have any roots inside the unit disc, then the ARMA process $\{\varepsilon_t\}$ is stationary and causal. Specifically, it has the representation $\varepsilon_t = \sum_{i=0}^{\infty} c_i \eta_{t-i}$ with $|c_i| = O(\rho^i)$ for some $\rho \in (0,1)$, implying that (C1)–(C3) are fulfilled. The results in Wu and Shao (2004) show that condition (C3) (as well as the other two conditions) is not only fulfilled for linear time series processes but also for a variety of non-linear processes.

3 The multiscale test

In this section, we introduce our multiscale method to test for local increases/decreases of the trend function m and analyse its theoretical properties. We assume throughout that m is continuously differentiable on $[0,1]$. The test problem under consideration can be formulated as follows: Let $H_0(u,h)$ be the hypothesis that m is constant on the interval $[u-h, u+h]$. Since m is differentiable, $H_0(u,h)$ can be reformulated as

$$H_0(u,h) : m'(w) = 0 \text{ for all } w \in [u-h, u+h],$$

where m' is the first derivative of m. We want to test the hypothesis $H_0(u,h)$ not only for a single interval $[u-h, u+h]$ but simultaneously for many different intervals. The overall null hypothesis is thus given by

$$H_0 : \text{ The hypothesis } H_0(u,h) \text{ holds true for all } (u,h) \in \mathcal{G}_T,$$

where $\mathcal{G}_T$ is some large set of points (u,h). The details on the set $\mathcal{G}_T$ are discussed at the end of Section 3.1 below. Note that $\mathcal{G}_T$ in general depends on the sample size T, implying that the null hypothesis $H_0 = H_{0,T}$ depends on T as well. We thus consider a sequence of null hypotheses $\{H_{0,T} : T = 1, 2, \ldots\}$ as T increases. For simplicity of notation, we however suppress the dependence of H_0 on T. In Sections 3.1 and 3.2, we step by step construct the multiscale test of the hypothesis H_0. The theoretical properties of the test are analysed in Section 3.3.

3.1 Construction of the multiscale statistic

We first construct a test statistic for the hypothesis $H_0(u,h)$, where $[u-h, u+h]$ is a given interval. To do so, we consider the kernel average

$$\widehat{\psi}_T(u,h) = \sum_{t=1}^{T} w_{t,T}(u,h) Y_{t,T},$$

where $w_{t,T}(u,h)$ is a kernel weight and h is the bandwidth. In order to avoid boundary issues, we work with a local linear weighting scheme. We in particular set

$$w_{t,T}(u,h) = \frac{\Lambda_{t,T}(u,h)}{\{\sum_{t=1}^{T} \Lambda_{t,T}(u,h)^2\}^{1/2}}, \tag{3.1}$$

where

$$\Lambda_{t,T}(u,h) = K\Big(\frac{\frac{t}{T}-u}{h}\Big)\Big[S_{T,0}(u,h)\Big(\frac{\frac{t}{T}-u}{h}\Big) - S_{T,1}(u,h)\Big],$$

$S_{T,\ell}(u,h) = (Th)^{-1}\sum_{t=1}^{T} K(\frac{\frac{t}{T}-u}{h})(\frac{\frac{t}{T}-u}{h})^{\ell}$ for $\ell = 0,1,2$ and K is a kernel function with the following properties:

(C4) The kernel K is non-negative, symmetric about zero and integrates to one. Moreover, it has compact support $[-1,1]$ and is Lipschitz continuous, that is, $|K(v) - K(w)| \leq C|v-w|$ for any $v, w \in \mathbb{R}$ and some constant $C > 0$.

The kernel average $\widehat{\psi}_T(u,h)$ is nothing else than a rescaled local linear estimator of the derivative $m'(u)$ with bandwidth h.[1]

A test statistic for the hypothesis $H_0(u,h)$ is given by the normalized kernel average $\widehat{\psi}_T(u,h)/\widehat{\sigma}$, where $\widehat{\sigma}^2$ is an estimator of the long-run variance $\sigma^2 = \sum_{\ell=-\infty}^{\infty} \mathrm{Cov}(\varepsilon_0, \varepsilon_\ell)$ of the error process $\{\varepsilon_t\}$. The problem of estimating σ^2 is discussed in detail in Section 4. For the time being, we suppose that $\widehat{\sigma}^2$ is an estimator with reasonable theoretical properties. Specifically, we assume that $\widehat{\sigma}^2 = \sigma^2 + o_p(\rho_T)$ with $\rho_T = o(1/\log T)$. This is a fairly weak condition which is in particular satisfied by the estimators of σ^2 analysed in Section 4. The kernel weights $w_{t,T}(u,h)$ are chosen such that in the case of independent errors ε_t, $\mathrm{Var}(\widehat{\psi}_T(u,h)) = \sigma^2$ for any location u and bandwidth h, where the long-run error variance σ^2 simplifies to $\sigma^2 = \mathrm{Var}(\varepsilon_t)$. In the more general case that the error terms satisfy the weak dependence conditions from Section 2, it holds that $\mathrm{Var}(\widehat{\psi}_T(u,h)) = \sigma^2 + o(1)$ for any location u and any bandwidth h with $h \to 0$ and $Th \to \infty$. Hence, for sufficiently large sample sizes T, the test statistic $\widehat{\psi}_T(u,h)/\widehat{\sigma}$ has approximately unit variance for any such u and h.

[1] Alternatively to the local linear weights defined in (3.1), we could also work with the weights $w_{t,T}(u,h) = K'(h^{-1}[u-t/T])/\{\sum_{t=1}^{T} K'(h^{-1}[u-t/T])^2\}^{1/2}$, where the kernel function K is assumed to be differentiable and K' is its derivative. We however prefer to use local linear weights as these have superior theoretical properties at the boundary.

We now combine the test statistics $\widehat{\psi}_T(u,h)/\widehat{\sigma}$ for a wide range of different locations u and bandwidths or scales h. There are different ways to do so, leading to different types of multiscale statistics. Our multiscale statistic is defined as

$$\widehat{\Psi}_T = \max_{(u,h)\in\mathcal{G}_T} \Big\{ \Big| \frac{\widehat{\psi}_T(u,h)}{\widehat{\sigma}} \Big| - \lambda(h) \Big\}, \tag{3.2}$$

where $\lambda(h) = \sqrt{2\log\{1/(2h)\}}$ and $\mathcal{G}_T$ is the set of points (u,h) that are taken into consideration. The details on the set $\mathcal{G}_T$ are given below. As can be seen, the statistic $\widehat{\Psi}_T$ does not simply aggregate the individual statistics $\widehat{\psi}_T(u,h)/\widehat{\sigma}$ by taking the supremum over all points $(u,h) \in \mathcal{G}_T$ as in more traditional multiscale approaches. We rather calibrate the statistics $\widehat{\psi}_T(u,h)/\widehat{\sigma}$ that correspond to the bandwidth h by subtracting the additive correction term $\lambda(h)$. This approach was pioneered by Dümbgen and Spokoiny (2001) and has been used in numerous other studies since then; see e.g. Dümbgen (2002), Rohde (2008), Dümbgen and Walther (2008), Rufibach and Walther (2010), Schmidt-Hieber et al. (2013) and Eckle et al. (2017). To see the heuristic idea behind the additive correction $\lambda(h)$, consider for a moment the uncorrected statistic

$$\widehat{\Psi}_{T,\text{uncorrected}} = \max_{(u,h)\in\mathcal{G}_T} \Big| \frac{\widehat{\psi}_T(u,h)}{\widehat{\sigma}} \Big|$$

and suppose that the hypothesis $H_0(u,h)$ is true for all $(u,h) \in \mathcal{G}_T$. For simplicity, assume that the errors ε_t are i.i.d. normally distributed and neglect the estimation error in $\widehat{\sigma}$, that is, set $\widehat{\sigma} = \sigma$. Moreover, suppose that the set $\mathcal{G}_T$ only consists of the points $(u_k, h_\ell) = ((2k-1)h_\ell, h_\ell)$ with $k = 1, \ldots, \lfloor 1/2h_\ell \rfloor$ and $\ell = 1, \ldots, L$. In this case, we can write

$$\widehat{\Psi}_{T,\text{uncorrected}} = \max_{1\le\ell\le L} \max_{1\le k\le\lfloor 1/2h_\ell\rfloor} \Big| \frac{\widehat{\psi}_T(u_k,h_\ell)}{\sigma} \Big|.$$

Under our simplifying assumptions, the statistics $\widehat{\psi}_T(u_k,h_\ell)/\sigma$ with $k = 1, \ldots, \lfloor 1/2h_\ell \rfloor$ are independent and standard normal for any given bandwidth h_ℓ. Since the maximum over $\lfloor 1/2h \rfloor$ independent standard normal random variables is $\lambda(h) + o_p(1)$ as $h \to 0$, we obtain that $\max_k \widehat{\psi}_T(u_k,h_\ell)/\sigma$ is approximately of size $\lambda(h_\ell)$ for small bandwidths h_ℓ. As $\lambda(h) \to \infty$ for $h \to 0$, this implies that $\max_k \widehat{\psi}_T(u_k,h_\ell)/\sigma$ tends to be much larger in size for small than for large bandwidths h_ℓ. As a result, the stochastic behaviour of the uncorrected statistic $\widehat{\Psi}_{T,\text{uncorrected}}$ tends to be dominated by the statistics $\widehat{\psi}_T(u_k,h_\ell)$ corresponding to small bandwidths h_ℓ. The additively corrected statistic $\widehat{\Psi}_T$, in contrast, puts the statistics $\widehat{\psi}_T(u_k,h_\ell)$ corresponding to different bandwidths h_ℓ on a more equal footing, thus counteracting the dominance of small bandwidth values.

The multiscale statistic $\widehat{\Psi}_T$ simultaneously takes into account all locations u and bandwidths h with $(u,h) \in \mathcal{G}_T$. Throughout the paper, we suppose that $\mathcal{G}_T$ is some subset of $\mathcal{G}_T^{\text{full}} = \{(u,h) : u = t/T \text{ for some } 1 \le t \le T \text{ and } h \in [h_{\min}, h_{\max}]\}$, where $h_{\min}$ and

$h_{\max}$ denote some minimal and maximal bandwidth value, respectively. For our theory to work, we require the following conditions to hold:

(C5) $|\mathcal{G}_T| = O(T^\theta)$ for some arbitrarily large but fixed constant $\theta > 0$, where $|\mathcal{G}_T|$ denotes the cardinality of $\mathcal{G}_T$.

(C6) $h_{\min} \gg T^{-(1-\frac{2}{q})} \log T$, that is, $h_{\min}/\{T^{-(1-\frac{2}{q})} \log T\} \to \infty$ with $q > 4$ defined in (C2) and $h_{\max} \leq 1/2$.

According to (C5), the number of points (u, h) in $\mathcal{G}_T$ should not grow faster than T^θ for some arbitrarily large but fixed $\theta > 0$. This is a fairly weak restriction as it allows the set $\mathcal{G}_T$ to be extremely large as compared to the sample size T. For example, we may work with the set

$$\mathcal{G}_T = \big\{(u, h) : u = t/T \text{ for some } 1 \leq t \leq T \text{ and } h \in [h_{\min}, h_{\max}] \\ \text{with } h = t/T \text{ for some } 1 \leq t \leq T\big\},$$

which contains more than enough points (u, h) for most practical applications. Condition (C6) imposes some restrictions on the minimal and maximal bandwidths $h_{\min}$ and $h_{\max}$. These conditions are fairly weak, allowing us to choose the bandwidth window $[h_{\min}, h_{\max}]$ extremely large. The lower bound on $h_{\min}$ depends on the parameter q defined in (C2) which specifies the number of existing moments for the error terms ε_t. As one can see, we can choose $h_{\min}$ to be of the order $T^{-1/2}$ for any $q > 4$. Hence, we can let $h_{\min}$ converge to 0 very quickly even if only the first few moments of the error terms ε_t exist. If all moments exist (i.e. $q = \infty$), we even get that $h_{\min}$ may converge to 0 almost as quickly as $T^{-1} \log T$. Moreover, the maximal bandwidth $h_{\max}$ is not even required to converge to 0, which implies that we can pick it very large.

Remark 3.1. *The above construction of the multiscale statistic can be easily adapted to hypotheses other than H_0. To do so, one simply needs to replace the kernel weights $w_{t,T}(u, h)$ defined in* (3.1) *by appropriate versions which are suited to test the hypothesis of interest. For example, if one wants to test for local convexity/concavity of m, one may define the kernel weights $w_{t,T}(u, h)$ such that the kernel average $\widehat{\psi}_T(u, h)$ is a (rescaled) estimator of the second derivative of m at the location u with bandwidth h.*

3.2 The test procedure

In order to formulate a test for the null hypothesis H_0, we still need to specify a critical value. To do so, we define the statistic

$$\Phi_T = \max_{(u,h)\in\mathcal{G}_T} \Big\{\Big|\frac{\phi_T(u, h)}{\sigma}\Big| - \lambda(h)\Big\}, \tag{3.3}$$

where $\phi_T(u,h) = \sum_{t=1}^T w_{t,T}(u,h)\,\sigma Z_t$ and Z_t are independent standard normal random variables. The statistic Φ_T can be regarded as a Gaussian version of the test statistic $\widehat{\Psi}_T$ under the null hypothesis H_0. Let $q_T(\alpha)$ be the $(1-\alpha)$-quantile of Φ_T. Importantly, the quantile $q_T(\alpha)$ can be computed by Monte Carlo simulations and can thus be regarded as known. Our multiscale test of the hypothesis H_0 is now defined as follows: For a given significance level $\alpha \in (0,1)$, we reject H_0 if $\widehat{\Psi}_T > q_T(\alpha)$.

3.3 Theoretical properties of the test

In order to examine the theoretical properties of our multiscale test, we introduce the auxiliary multiscale statistic

$$\widehat{\Phi}_T = \max_{(u,h)\in\mathcal{G}_T} \Big\{ \Big| \frac{\widehat{\phi}_T(u,h)}{\widehat{\sigma}} \Big| - \lambda(h) \Big\} \tag{3.4}$$

with $\widehat{\phi}_T(u,h) = \widehat{\psi}_T(u,h) - \mathbb{E}[\widehat{\psi}_T(u,h)] = \sum_{t=1}^T w_{t,T}(u,h)\varepsilon_t$. The following result is central to the theoretical analysis of our multiscale test. According to it, the (known) quantile $q_T(\alpha)$ of the Gaussian statistic Φ_T defined in Section 3.2 can be used as a proxy for the $(1-\alpha)$-quantile of the multiscale statistic $\widehat{\Phi}_T$.

Theorem 3.1. *Let (C1)–(C6) be fulfilled and assume that* $\widehat{\sigma}^2 = \sigma^2 + o_p(\rho_T)$ *with* $\rho_T = o(1/\log T)$. *Then*

$$\mathbb{P}\big(\widehat{\Phi}_T \le q_T(\alpha)\big) = (1-\alpha) + o(1).$$

A full proof of Theorem 3.1 is given in the Appendix. We here shortly outline the proof strategy, which splits up into two main steps. In the first, we replace the statistic $\widehat{\Phi}_T$ for each $T \ge 1$ by a statistic $\widetilde{\Phi}_T$ with the same distribution as $\widehat{\Phi}_T$ and the property that

$$\big|\widetilde{\Phi}_T - \Phi_T\big| = o_p(\delta_T), \tag{3.5}$$

where $\delta_T = o(1)$ and the Gaussian statistic Φ_T is defined in Section 3.2. We thus replace the statistic $\widehat{\Phi}_T$ by an identically distributed version which is close to a Gaussian statistic whose distribution is known. To do so, we make use of strong approximation theory for dependent processes as derived in Berkes et al. (2014). In the second step, we show that

$$\sup_{x\in\mathbb{R}} \big|\mathbb{P}(\widetilde{\Phi}_T \le x) - \mathbb{P}(\Phi_T \le x)\big| = o(1), \tag{3.6}$$

which immediately implies the statement of Theorem 3.1. Importantly, the convergence result (3.5) is not sufficient for establishing (3.6). Put differently, the fact that $\widetilde{\Phi}_T$ can be approximated by Φ_T in the sense that $\widetilde{\Phi}_T - \Phi_T = o_p(\delta_T)$ does not imply that the distribution of $\widetilde{\Phi}_T$ is close to that of Φ_T in the sense of (3.6). For (3.6) to hold, we

additionally require the distribution of Φ_T to have some sort of continuity property. Specifically, we prove that

$$\sup_{x\in\mathbb{R}} \mathbb{P}\big(|\Phi_T - x| \leq \delta_T\big) = o(1), \tag{3.7}$$

which says that Φ_T does not concentrate too strongly in small regions of the form $[x-\delta_T, x+\delta_T]$. The main tool for verifying (3.7) are anti-concentration results for Gaussian random vectors as derived in Chernozhukov et al. (2015). The claim (3.6) can be proven by using (3.5) together with (3.7), which in turn yields Theorem 3.1.

The main idea of our proof strategy is to combine strong approximation theory with anti-concentration bounds for Gaussian random vectors to show that the quantiles of the multiscale statistic $\widehat{\Phi}_T$ can be proxied by those of a Gaussian analogue. This strategy is quite general in nature and may be applied to other multiscale problems for dependent data. Strong approximation theory has also been used to investigate multiscale tests for independent data; see e.g. Schmidt-Hieber et al. (2013). However, it has not been combined with anti-concentration results to approximate the quantiles of the multiscale statistic. As an alternative to strong approximation theory, Eckle et al. (2017) and Proksch et al. (2018) have recently used Gaussian approximation results derived in Chernozhukov et al. (2014, 2017) to analyse multiscale tests for independent data. Even though it might be possible to adapt these techniques to the case of dependent data, this is not trivial at all as part of the technical arguments and the Gaussian approximation tools strongly rely on the assumption of independence.

We now investigate the theoretical properties of our multiscale test with the help of Theorem 3.1. The first result is an immediate consequence of Theorem 3.1. It says that the test has the correct (asymptotic) size.

Proposition 3.1. *Let the conditions of Theorem 3.1 be satisfied. Under the null hypothesis H_0, it holds that*

$$\mathbb{P}\big(\widehat{\Psi}_T \leq q_T(\alpha)\big) = (1-\alpha) + o(1).$$

The second result characterizes the power of the multiscale test against local alternatives. To formulate it, we consider any sequence of functions $m = m_T$ with the following property: There exists $(u,h) \in \mathcal{G}_T$ with $[u-h, u+h] \subseteq [0,1]$ such that

$$m_T'(w) \geq c_T \sqrt{\frac{\log T}{Th^3}} \quad \text{for all } w \in [u-h, u+h], \tag{3.8}$$

where $\{c_T\}$ is any sequence of positive numbers with $c_T \to \infty$. Alternatively to (3.8), we may also assume that $-m_T'(w) \geq c_T\sqrt{\log T/(Th^3)}$ for all $w \in [u-h, u+h]$. According to the following result, our test has asymptotic power 1 against local alternatives of the form (3.8).

Proposition 3.2. *Let the conditions of Theorem 3.1 be satisfied and consider any sequence of functions m_T with the property* (3.8)*. Then*

$$\mathbb{P}\big(\widehat{\Psi}_T \leq q_T(\alpha)\big) = o(1).$$

The proof of Proposition 3.2 can be found in the Appendix. To formulate the next result, we define

$$\begin{aligned}
\Pi_T^{\pm} &= \big\{I_{u,h} = [u-h, u+h] : (u,h) \in \mathcal{A}_T^{\pm}\big\} \\
\Pi_T^{+} &= \big\{I_{u,h} = [u-h, u+h] : (u,h) \in \mathcal{A}_T^{+} \text{ and } I_{u,h} \subseteq [0,1]\big\} \\
\Pi_T^{-} &= \big\{I_{u,h} = [u-h, u+h] : (u,h) \in \mathcal{A}_T^{-} \text{ and } I_{u,h} \subseteq [0,1]\big\}
\end{aligned}$$

together with

$$\begin{aligned}
\mathcal{A}_T^{\pm} &= \Big\{(u,h) \in \mathcal{G}_T : \Big|\frac{\widehat{\psi}_T(u,h)}{\widehat{\sigma}}\Big| > q_T(\alpha) + \lambda(h)\Big\} \\
\mathcal{A}_T^{+} &= \Big\{(u,h) \in \mathcal{G}_T : \frac{\widehat{\psi}_T(u,h)}{\widehat{\sigma}} > q_T(\alpha) + \lambda(h)\Big\} \\
\mathcal{A}_T^{-} &= \Big\{(u,h) \in \mathcal{G}_T : -\frac{\widehat{\psi}_T(u,h)}{\widehat{\sigma}} > q_T(\alpha) + \lambda(h)\Big\}.
\end{aligned}$$

$\Pi_T^{\pm}$ is the collection of intervals $I_{u,h} = [u-h, u+h]$ for which the (corrected) test statistic $|\widehat{\psi}_T(u,h)/\widehat{\sigma}| - \lambda(h)$ lies above the critical value $q_T(\alpha)$. Π_T^{+} and Π_T^{-} can be interpreted analogously but take into account the sign of the statistic $\widehat{\psi}_T(u,h)/\widehat{\sigma}$. With this notation at hand, we consider the events

$$\begin{aligned}
E_T^{\pm} &= \Big\{\forall I_{u,h} \in \Pi_T^{\pm} : m'(v) \neq 0 \text{ for some } v \in I_{u,h} = [u-h, u+h]\Big\} \\
E_T^{+} &= \Big\{\forall I_{u,h} \in \Pi_T^{+} : m'(v) > 0 \text{ for some } v \in I_{u,h} = [u-h, u+h]\Big\} \\
E_T^{-} &= \Big\{\forall I_{u,h} \in \Pi_T^{-} : m'(v) < 0 \text{ for some } v \in I_{u,h} = [u-h, u+h]\Big\}.
\end{aligned}$$

$E_T^{\pm}$ (E_T^{+}, E_T^{-}) is the event that the function m is non-constant (increasing, decreasing) on all intervals $I_{u,h} \in \Pi_T^{\pm}$ (Π_T^{+}, Π_T^{-}). More precisely, $E_T^{\pm}$ (E_T^{+}, E_T^{-}) is the event that for each interval $I_{u,h} \in \Pi_T^{\pm}$ (Π_T^{+}, Π_T^{-}), there is a subset $J_{u,h} \subseteq I_{u,h}$ with m being a non-constant (increasing, decreasing) function on $J_{u,h}$. We can make the following formal statement about the events $E_T^{\pm}$, E_T^{+} and E_T^{-} whose proof is given in the Appendix.

Proposition 3.3. *Let the conditions of Theorem 3.1 be fulfilled. Then for $\ell \in \{\pm, +, -\}$, it holds that*

$$\mathbb{P}\big(E_T^{\ell}\big) \geq (1-\alpha) + o(1).$$

According to Proposition 3.3, we can make uniform confidence statements of the following form: With (asymptotic) probability $\geq (1-\alpha)$, the trend function m is non-constant

(increasing, decreasing) on some part of the interval $I_{u,h}$ for all $I_{u,h} \in \Pi_T$ (Π_T^+, Π_T^-). Hence, our multiscale procedure allows to identify, with a pre-specified confidence, time regions where there is an increase/decrease in the time trend m.

Remark 3.2. *Unlike Π_T, the sets Π_T^+ and Π_T^- only contain intervals $I_{u,h} = [u-h, u+h]$ which are subsets of $[0,1]$. We thus exclude points $(u,h) \in \mathcal{A}_T^+$ and $(u,h) \in \mathcal{A}_T^-$ which lie at the boundary, that is, for which $I_{u,h} \not\subseteq [0,1]$. The reason is as follows: Let $(u,h) \in \mathcal{A}_T^+$ with $I_{u,h} \not\subseteq [0,1]$. Our technical arguments allow us to say, with asymptotic confidence $\geq 1-\alpha$, that $m'(v) \neq 0$ for some $v \in I_{u,h}$. However, we cannot say whether $m'(v) > 0$ or $m'(v) < 0$, that is, we cannot make confidence statements about the sign. Crudely speaking, the problem is that the local linear weights $w_{t,T}(u,h)$ behave quite differently at boundary points (u,h) with $I_{u,h} \not\subseteq [0,1]$. As a consequence, we can include boundary points (u,h) in Π_T but not in Π_T^+ and Π_T^-.*

The statement of Proposition 3.3 suggests to graphically present the results of our multiscale test by plotting the intervals $I_{u,h} \in \Pi_T^\ell$ for $\ell \in \{\pm, +, -\}$, that is, by plotting the intervals where (with asymptotic confidence $\geq 1-\alpha$) our test detects a violation of the null hypothesis. The drawback of this graphical presentation is that the number of intervals in Π_T^ℓ is often quite large. To obtain a better graphical summary of the results, we replace Π_T^ℓ by a subset $\Pi_T^{\ell,\min}$ which is constructed as follows: As in Dümbgen (2002), we call an interval $I_{u,h} \in \Pi_T^\ell$ minimal if there is no other interval $I_{u',h'} \in \Pi_T^\ell$ with $I_{u',h'} \subset I_{u,h}$. Let $\Pi_T^{\ell,\min}$ be the set of all minimal intervals in Π_T^ℓ for $\ell \in \{\pm, +, -\}$ and define the events

$$
\begin{aligned}
E_T^{\pm,\min} &= \Big\{ \forall I_{u,h} \in \Pi_T^{\pm,\min} : m'(v) \neq 0 \text{ for some } v \in I_{u,h} = [u-h, u+h] \Big\} \\
E_T^{+,\min} &= \Big\{ \forall I_{u,h} \in \Pi_T^{+,\min} : m'(v) > 0 \text{ for some } v \in I_{u,h} = [u-h, u+h] \Big\} \\
E_T^{-,\min} &= \Big\{ \forall I_{u,h} \in \Pi_T^{-,\min} : m'(v) < 0 \text{ for some } v \in I_{u,h} = [u-h, u+h] \Big\}.
\end{aligned}
$$

It is easily seen that $E_T^\ell = E_T^{\ell,\min}$ for $\ell \in \{\pm, +, -\}$. Hence, by Proposition 3.3, it holds that

$$
\mathbb{P}\big(E_T^{\ell,\min}\big) \geq (1-\alpha) + o(1)
$$

for $\ell \in \{\pm, +, -\}$. This suggests to plot the minimal intervals in $\Pi_T^{\ell,\min}$ rather than the whole collection of intervals Π_T^ℓ as a graphical summary of the test results. We in particular use this way of presenting the test results in our applications of Section **??**.

4 Estimation of the long-run error variance

We now discuss how to estimate the long-run error variance $\sigma^2 = \sum_{\ell=-\infty}^{\infty} \mathrm{Cov}(\varepsilon_0, \varepsilon_\ell)$ in model (2.1). There are two broad classes of estimators: residual- and difference-based

estimators. Residual-based methods proceed by applying standard methods for estimating σ^2 to the time series of residuals $\widehat{\varepsilon}_t = Y_{t,T} - \widehat{m}_h(t/T)$, where $\widehat{m}_h$ is a nonparametric estimator of m with the bandwidth or smoothing parameter h. Difference-based methods attempt to estimate σ^2 by applying statistical methods to certain differences of the observed time series $\{Y_{t,T}\}$, for example, to the first differences $\Delta Y_{t,T} = Y_{t,T} - Y_{t-1,T}$. Notably, difference-based methods do not involve a nonparametric estimator of m and thus do not require to specify a smoothing parameter h for the estimation of m.

4.1 General weakly dependent error terms

We first consider a general stationary error process $\{\varepsilon_t\}$. In particular, we do not impose any time series model such as a moving average (MA) or an autoregressive (AR) model on $\{\varepsilon_t\}$ but only require that $\{\varepsilon_t\}$ satisfies certain weak dependence conditions such as those from Section 2.

A residual-based estimator of σ^2 can be obtained as follows: Estimating σ^2 amounts to estimating the spectral density f_ε of the error process $\{\varepsilon_t\}$ at frequency 0 (assuming that f_ε exists). We may thus apply existing methods for estimating $f_\varepsilon(0)$ such as developed in Liu and Wu (2010) to the time series of residuals $\widehat{\varepsilon}_t = Y_{t,T} - \widehat{m}_h(t/T)$. Note that the resulting estimator of σ^2 will depend on two different smoothing parameters: one for the preliminary estimation of m and one for the estimation of the nonparametric density f_ε at the point 0. Alternatively, σ^2 may be estimated by difference-based methods. We briefly describe an approach which goes back to ideas from Hart (1989, 1991). Consider the differences $\Delta Y_{t,T} = Y_{t,T} - Y_{t-1,T}$ and let I_Δ be (a tapered version of) the periodogram of $\{\Delta_t\}$.[2] Following Hart (1989, 1991), the autocovariances $\gamma_\varepsilon(\ell) = \mathrm{Cov}(\varepsilon_0, \varepsilon_\ell)$ can be estimated by

$$\widehat{\gamma}_\varepsilon(\ell) = \frac{4\pi}{T} \sum_{j=j(\delta)}^{\lceil T/2 \rceil} \cos(\omega_j \ell) |1 - \exp(-i\omega_j)|^{-2} I_\Delta(\omega_j),$$

where $\omega_j = 2\pi j/T$, $j(\delta) = \lceil n\delta/(2\pi) \rceil$ and δ is a tuning parameter satisfying $\delta \to 0$ and $T\delta \to \infty$. We may now employ HAC-type estimation procedures, as discussed in Andrews (1991) or De Jong and Davidson (2000), to estimate σ^2 by

$$\widehat{\sigma}^2 = \sum_{|\ell| \le b_T} W\Big(\frac{\ell}{b_T}\Big) \widehat{\gamma}_\varepsilon(\ell),$$

where $W : [-1, 1] \to \mathbb{R}$ is a kernel of Bartlett or flat-top type and b_T is a bandwidth parameter with $b_T \to \infty$ and $b_T/T \to 0$.

Estimating the long-run error variance of $\{\varepsilon_t\}$ under general conditions is a notoriously difficult problem and estimators of σ^2 such as those described above tend to be quite

[2] Hart (1989, 1991) actually considered the differences $\Delta_2 Y_{t,T} = Y_{t+1,T} - 2Y_{t,T} + Y_{t-1,T}$. However, the idea behind the estimation method is the same no matter whether $\Delta Y_{t,T}$ or $\Delta_2 Y_{t,T}$ is used.

imprecise. Moreover, no matter whether residual- or difference-based methods are used, the estimators depend on one or several smoothing parameters which are quite difficult to select. For these reasons, we follow authors such as Hart (1991, 1994) and Hall and Van Keilegom (2003) and impose a time series model on the error terms $\{\varepsilon_t\}$ in model (2.1). Estimating σ^2 under the restrictions of such a model may of course create some misspecification bias. However, as long as the model gives a reasonable approximation to the true error process, the produced estimates of σ^2 can be expected to be fairly reliable even though they are a bit biased.

4.2 Autoregressive error terms

A number of studies have analysed the problem of estimating σ^2 in model (2.1) with MA(q) or, more generally, q-dependent error terms. Difference-based estimators of σ^2 for this case have been proposed in Müller and Stadtmüller (1988), Herrmann et al. (1992) and Tecuapetla-Gómez and Munk (2017) among others. Under the assumption of q-dependence, $\gamma_\varepsilon(\ell) := \mathrm{Cov}(\varepsilon_0, \varepsilon_\ell) = 0$ for all $|\ell| > q$. Even though q-dependent time series are a reasonable error model in some applications, the condition that $\gamma_\varepsilon(\ell)$ is exactly equal to 0 for sufficiently large lags ℓ is quite restrictive in many situations. Presumably the most widely used error model in practice is an AR(p) process. Residual-based methods to estimate σ^2 in model (2.1) with AR(p) errors can be found for example in Truong (1991), Shao and Yang (2011) and Qiu et al. (2013). A difference-based method was proposed in Hall and Van Keilegom (2003). Even though this method has the advantage of not involving any bandwidth parameter for the estimation of the trend function m, it depends on some other tuning parameters.

In what follows, we introduce an estimator of σ^2 for the AR(p) case which is completely free of tuning parameters and very simple to compute. In particular, it does not involve any numerical optimization but can be computed in closed form. As we will see, the proposed estimation method can also be extended to ARMA(p, q) errors in a straightforward way. As in Hall and Van Keilegom (2003), we consider the following situation: $\{\varepsilon_t\}$ is a stationary and causal AR(p) process of the form

$$\varepsilon_t = \sum_{j=1}^{p} a_j^* \varepsilon_{t-j} + \eta_t, \tag{4.1}$$

where $a_1^*, \ldots, a_p^*$ are unknown parameters and η_t are i.i.d. innovations with $\mathbb{E}[\eta_t] = 0$ and $\mathbb{E}[\eta_t^2] = \nu^*$. The AR order p is known and m is Lipschitz continuous on $[0, 1]$, that is, $|m(u) - m(v)| \leq L|u - v|$ for all $u, v \in [0, 1]$ and some constant $L < \infty$.

Our estimation method relies on the following simple observation: If $\{\varepsilon_t\}$ is an AR(p) process of the form (4.1), then the time series $\{\Delta\varepsilon_t\}$ of the differences $\Delta\varepsilon_t = \varepsilon_t - \varepsilon_{t-1}$

is an ARMA$(p,1)$ process of the form

$$\Delta\varepsilon_t - \sum_{j=1}^{p} a_j^* \Delta\varepsilon_{t-j} = \eta_t - \eta_{t-1}. \tag{4.2}$$

The differences $\Delta\varepsilon_t$ of the unobserved error process are close to the differences $\Delta Y_{t,T} = Y_{t,T} - Y_{t-1,T}$ of the observed time series in the sense that

$$\Delta Y_{t,T} = \big[\varepsilon_t - \varepsilon_{t-1}\big] + \Big[m\Big(\frac{t}{T}\Big) - m\Big(\frac{t-1}{T}\Big)\Big] = \Delta\varepsilon_t + O\Big(\frac{1}{T}\Big). \tag{4.3}$$

Taken together, (4.2) and (4.3) imply that the differenced time series $\{\Delta Y_{t,T}\}$ is approximately an ARMA$(p,1)$ process of the form (4.2). This suggests to estimate the model parameters $\boldsymbol{a}^* = (a_1^*, \ldots, a_p^*)$ and ν^* by applying estimation methods for ARMA processes to the time series $\{\Delta Y_{t,T}\}$. Since $\sigma^2 = \nu^*(1 - \sum_{j=1}^{p} a_j^*)^{-2}$ in the AR(p) case under consideration, we immediately obtain an estimator of σ^2 as well. Notably, the ARMA process $\{\Delta\varepsilon_t\}$ defined in (4.2) is non-standard in the sense that the MA(1) polynomial $B(z) = 1 - z$ has a unit root, implying that the process is not invertible. Maximum likelihood methods for ARMA process with unit roots in the MA polynomial were for example developed in Pham-Dinh (1978). The class of ARMA models considered there nests the process $\{\Delta\varepsilon_t\}$ as a special case. To construct estimators of the model parameters $(\boldsymbol{a}^*, \nu^*)$ and the long-run variance σ^2, we apply the maximum likelihood approach of Pham-Dinh (1978) to the differences $\Delta Y_{t,T}$ which are contaminated by the smooth trend m. Our theoretical analysis will show that the resulting estimators of $\boldsymbol{a}^*$, ν^* and σ^2 have the same asymptotic properties as those based on the unobserved ARMA process $\{\Delta\varepsilon_t\}$, that is, the influence of the smooth trend m on the estimators can be neglected asymptotically.

To define our estimators of $\boldsymbol{a}^*$, ν^* and σ^2, we introduce the random variables

$$Q_{t,j} = \sum_{\ell=p+2}^{t} \Delta Y_{\ell-j,T} - \frac{1}{t} \sum_{s=p+2}^{t-1} \sum_{\ell=p+2}^{s} \Delta Y_{\ell-j,T}$$

for $t \geq p+2$ and $0 \leq j \leq p$ together with $\widehat{\gamma}_Q(i,j) = \sum_{t=p+2}^{T} Q_{t,i} Q_{t,j}$. With this notation at hand, our estimators of $\boldsymbol{a}^*$ and ν^* are given by

$$\widehat{\boldsymbol{a}} = \widehat{\boldsymbol{\Gamma}}_Q^{-1} \widehat{\boldsymbol{\gamma}}_Q \qquad \text{and} \qquad \widehat{\nu} = \frac{1}{T-p-1} \sum_{t=p+2}^{T} \Big(Q_{t,0} - \sum_{j=1}^{p} \widehat{a}_j Q_{t,j}\Big)^2, \tag{4.4}$$

where $\widehat{\boldsymbol{\Gamma}}_Q = (\widehat{\gamma}_Q(i,j) : 1 \leq i,j \leq p)$ and $\widehat{\boldsymbol{\gamma}}_Q = (\widehat{\gamma}_Q(0,1), \ldots, \widehat{\gamma}_Q(0,p))^\top$. Moreover, the

long-run error variance σ^2 is estimated by

$$\widehat{\sigma}^2 = \widehat{\nu}\Big(1 - \sum_{j=1}^{p} \widehat{a}_j\Big)^{-2}. \tag{4.5}$$

The idea behind the estimators $\widehat{\boldsymbol{a}}$ and $\widehat{\nu}$ is as follows: Suppose that the differences $\Delta\varepsilon_t$ were observed for $1 \leq t \leq T$ and let $\Pi_s(\Delta\varepsilon_t)$ be the orthogonal projection of $\Delta\varepsilon_t$ onto the linear space spanned by $\Delta\varepsilon_1, \ldots, \Delta\varepsilon_s$. The projection $\Pi_{t-1}(\Delta\varepsilon_t)$ is the best linear predictor of $\Delta\varepsilon_t$ based on $\Delta\varepsilon_1, \ldots, \Delta\varepsilon_{t-1}$. Denote the prediction innovations by $\xi_t(\boldsymbol{a}^*, \nu^*) = \Delta\varepsilon_t - \Pi_{t-1}(\Delta\varepsilon_t)$ and let $e_t(\boldsymbol{a}^*, \nu^*) = \mathbb{E}[\xi_t^2(\boldsymbol{a}^*, \nu^*)]$ be the corresponding prediction error. Under the assumption that the innovations $\xi_t(\boldsymbol{a}^*, \nu^*)$ are i.i.d. Gaussian, the log-likelihood is given by

$$\mathcal{L}_T(\boldsymbol{a}^*, \nu^*) = -\frac{1}{2}\sum_{t=1}^{T} \log\big(2\pi e_t(\boldsymbol{a}^*, \nu^*)\big) - \frac{1}{2}\sum_{t=1}^{T} \frac{\xi_t^2(\boldsymbol{a}^*, \nu^*)}{e_t(\boldsymbol{a}^*, \nu^*)}.$$

Following the arguments in Pham-Dinh (1978), the prediction innovations $\xi_t(\boldsymbol{a}^*, \nu^*)$ and the prediction errors $e_t(\boldsymbol{a}^*, \nu^*)$ can be approximated by $\widehat{\xi}_t(\boldsymbol{a}) = Q_{t,0} - \sum_{j=1}^{p} a_j Q_{t,j}$ and ν^* in the likelihood function $\mathcal{L}_T(\boldsymbol{a}^*, \nu^*)$. Plugging these terms into $\mathcal{L}_T(\boldsymbol{a}^*, \nu^*)$ (and taking into account that they can only be computed for $t > p+1$ in practice), we obtain the approximate log-likelihood

$$L_T(\boldsymbol{a}^*, \nu^*) = -\frac{T-p-1}{2}\log(2\pi\nu^*) - \frac{1}{2\nu^*}\sum_{t=p+2}^{T}\Big(Q_{t,0} - \sum_{j=1}^{p} a_j^* Q_{t,j}\Big)^2.$$

The model parameters $(\boldsymbol{a}^*, \nu^*)$ can now be estimated by the maximizers $(\widehat{\boldsymbol{a}}, \widehat{\nu}) = \arg\max_{\boldsymbol{a},\nu} L_T(\boldsymbol{a}, \nu)$ of the log-likelihood $L_T(\boldsymbol{a}, \nu)$. This maximization problem can be solved analytically and yields the solutions defined in (4.4). Our estimators have the following formal properties.

Proposition 4.1. *Let $\{\varepsilon_t\}$ be a stationary and causal AR(p) process of the form (4.1), suppose that the innovations η_t have a finite fourth moment, and let m be Lipschitz. Then $\widehat{\boldsymbol{a}} = \boldsymbol{a}^* + O_p(T^{-1/2})$, $\widehat{\nu} = \nu^* + O_p(T^{-1/2})$ and $\widehat{\sigma}^2 = \sigma^2 + O_p(T^{-1/2})$. Moreover,*

$$\sqrt{T}\left(\begin{pmatrix}\widehat{\boldsymbol{a}}\\ \widehat{\nu}\end{pmatrix} - \begin{pmatrix}\boldsymbol{a}^*\\ \nu^*\end{pmatrix}\right) \xrightarrow{d} N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\nu^*\Gamma^{-1} & 0\\ 0 & 2(\nu^*)^2 + \kappa\end{pmatrix}\right),$$

where $\boldsymbol{\Gamma}$ is the autocovariance matrix of the AR(p) process $\{\varepsilon_t\}$ and κ is the fourth cumulant of $\{\eta_t\}$.

According to Proposition ??, our estimators are $\sqrt{T}$-consistent. Moreover, as can be seen, the parameter estimators $\widehat{\boldsymbol{a}}$ are asymptotically normal with limit variance ??. This is the same variance as ??

We briefly compare our estimators to competing methods from the literature. Presumably closest to our method is the procedure of Hall and Van Keilegom (2003). Nevertheless, it differs from our approach in several respects, the most important difference being the following: In order to construct their estimators, Hall and Van Keilegom (2003) need to compute the differences $\Delta_\ell \varepsilon_t = \varepsilon_t - \varepsilon_{t-\ell}$ for $L_1 \leq \ell \leq L_2$, where L_1 and L_2 are two tuning parameters that are required to fulfill the conditions $L_1/\log T \to \infty$ and $L_2 = O(T^{1/2})$. Our approach, in contrast, is only based on the first differences $\Delta\varepsilon_t = \varepsilon_t - \varepsilon_{t-1}$ and does not involve any tuning parameters. The advantage of being completely free of tuning parameters comes, of course, at some cost: Our estimators of the AR parameters $a_1, \ldots, a_p$ (and thus also those of σ_η^2 and σ^2) are $\sqrt{T}$-consistent as are the estimators of Hall and Van Keilegom (2003). However, in general, they have a somewhat larger asymptotic variance than those of Hall and Van Keilegom (2003) and in this sense are a bit less efficient. Hence, we trade a bit of efficiency for the practical advantage of not having to choose any tuning parameter.

Before we close the section, we discuss how to extend our methods to the ARMA case. To do so, suppose that $\{\varepsilon_t\}$ is a stationary and causal ARMA(p, q) process of the form

$$\varepsilon_t - \sum_{j=1}^{p} a_j \varepsilon_{t-j} = \eta_t + \sum_{k=1}^{q} b_k \eta_{t-k}, \tag{4.6}$$

where $a_1, \ldots, a_p$ and $b_1, \ldots, b_q$ are unknown parameters and the innovations η_t are as above. Similarly as in the AR(p) case, we can show the following: If $\{\varepsilon_t\}$ is an ARMA(p, q) process of the form (4.6), then the time series $\{\Delta\varepsilon_t\}$ of the differences $\Delta\varepsilon_t = \varepsilon_t - \varepsilon_{t-1}$ is an ARMA$(p, q+1)$ process of the form

$$\Delta\varepsilon_t - \sum_{j=1}^{p} a_j \Delta\varepsilon_{t-j} = \eta_t + \sum_{k=1}^{q+1} d_k \eta_{t-k}, \tag{4.7}$$

where $d_k = b_k + b_{k-1}$ for $1 \leq k \leq q$ with $b_0 = 1$ and $d_{q+1} = b_q$. The unknown parameters $a_1, \ldots, a_p$, $d_1, \ldots, d_{q+1}$ and σ_η^2 can be estimated similarly as above by using the Yule-Walker equations corresponding to model (4.7). Once we have computed estimators for $d_1, \ldots, d_{q+1}$, we trivially get estimators for $b_1, \ldots, b_q$ as well and can estimate the long-run error variance σ^2 with the help of the formula $\sigma^2 = \sigma_\eta^2 (1 + \sum_{k=1}^{q} b_k)^2 / (1 - \sum_{j=1}^{p} a_j)^2$.

4.3 Parameter estimation in ARMA models with a unit root in the MA polynomial

Let $\{X_t\}$ be a stationary causal ARMA$(p,1)$ process of the form

$$X_t - \sum_{j=1}^{p} a_j^* X_{t-j} = \eta_t - \eta_{t-1}, \tag{4.8}$$

where η_t are i.i.d. innovations with $\mathbb{E}[\eta_t] = 0$ and $\mathbb{E}[\eta^2] = \nu^*$. We use the notation $\boldsymbol{a}^* = (a_1^*, \ldots, a_p^*)$ and $\boldsymbol{\theta}^* = (\boldsymbol{a}^*, \nu^*)$. We now construct a maximum likelihood estimator of the parameters $a_1^*, \ldots, a_p^*$ and ω^*. The constuction proceeds in two steps: We first define an infeasible likelihood function which cannot be computed in practice and then approximate it by a feasible version.

Step 1. Let $\Pi_s Z_t$ be the orthogonal projection of a general (square-integrable) random variable Z_t onto the linear space spanned by $X_1, \ldots, X_s$, denoted by $\text{span}\{X_1, \ldots, X_s\}$. The projection $\Pi_{t-1} X_t$ is the best linear predictor of X_t based on $X_1, \ldots, X_{t-1}$. Let $\xi_t(\boldsymbol{\theta}^*) = X_t - \Pi_{t-1} X_t$ be the prediction innovations and $e_t(\boldsymbol{\theta}^*) = \mathbb{E}[\xi_t^2(\boldsymbol{\theta}^*)]$ the corresponding prediction error. Under the assumption that the innovations $\xi_t(\boldsymbol{\theta}^*)$ are i.i.d. Gaussian, the (infeasible) log-likelihood is given by

$$\mathcal{L}_T(\boldsymbol{\theta}^*) = -\frac{1}{2}\sum_{t=1}^{T} \log\big(2\pi e_t(\boldsymbol{\theta}^*)\big) - \frac{1}{2}\sum_{t=1}^{T} \frac{\xi_t^2(\boldsymbol{\theta}^*)}{e_t(\boldsymbol{\theta}^*)}.$$

The prediction innovations $\xi_t(\boldsymbol{\theta}^*)$ and the prediction errors $e_t(\boldsymbol{\theta}^*)$ can be shown to have the following representations:

$$\xi_t(\boldsymbol{\theta}^*) = V_t(\boldsymbol{a}^*) - \frac{1}{\beta(\nu^*) + t}\sum_{s=p+1}^{t-1} V_s(\boldsymbol{a}^*) + \frac{\beta(\nu^*) + p + 1}{\beta(\nu^*) + t}\Pi_p \eta_p \tag{4.9}$$

$$e_t(\boldsymbol{\theta}^*) = \Big(1 + \frac{1}{\beta(\nu^*) + t}\Big)\nu^* \tag{4.10}$$

for $t > p$, where $V_t(\boldsymbol{a}^*) = \sum_{k=p+1}^{s}(X_t - \sum_{j=1}^{p} a_j^* X_{t-j})$ and $\beta(\nu^*) = (\nu^*/\mu_p) - p - 1$ with $\mu_p = \mathbb{E}[(\eta_p - \Pi_p \eta_p)^2]$.

Step 2. For a general parameter vector $\boldsymbol{\theta} = (\boldsymbol{a}, \nu) = (a_1, \ldots, a_p, \nu)$, we approximate the innovations $\xi_t(\boldsymbol{\theta})$ by

$$\widehat{\xi}_t(\boldsymbol{\theta}) = V_t(\boldsymbol{a}) - \frac{1}{t}\sum_{s=p+1}^{t-1} V_s(\boldsymbol{a})$$

and the prediction error $e_t(\boldsymbol{\theta}) = \mathbb{E}[\xi_t^2(\boldsymbol{\theta})]$ by ν. A more convenient representation of

$\widehat{\xi}_t(\boldsymbol{\theta})$ is given by

$$\widehat{\xi}_t(\boldsymbol{\theta}) = Q_{t,0} - \sum_{j=1}^{p} a_j Q_{t,j} \quad \text{with} \quad Q_{t,j} = \sum_{\ell=p+1}^{t} X_{\ell-j} - \frac{1}{t}\sum_{s=p+1}^{t-1}\sum_{\ell=p+1}^{s} X_{\ell-j}.$$

Replacing $\xi_t(\boldsymbol{\theta})$ and $e_t(\boldsymbol{\theta})$ by the approximations $\widehat{\xi}_t(\boldsymbol{\theta})$ and ν in in $\mathcal{L}_T(\boldsymbol{\theta})$ yields the feasible likelihood

$$L_T(\boldsymbol{\theta}) = -\frac{T-p}{2}\log(2\pi\nu) - \frac{1}{2\nu}\sum_{t=p+1}^{T}\widehat{\xi}_t^2(\boldsymbol{\theta}).$$

Estimators $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{a}}, \widehat{\nu})$ of the parameters $\boldsymbol{\theta}^* = (\boldsymbol{a}^*, \nu^*)$ are defined as

$$\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{a}}, \widehat{\nu}) = \arg\max_{\theta\in\Theta} L_T(\boldsymbol{\theta}).$$

It is straightforward to solve this maximization problem and to show that

$$\begin{aligned}
\widehat{\boldsymbol{a}} &= \widehat{\boldsymbol{\Gamma}}_Q^{-1}\widehat{\boldsymbol{\gamma}}_Q \\
\widehat{\nu} &= \frac{1}{T-p}\sum_{t=p+1}^{T}\Big(Q_{t,0} - \sum_{j=1}^{p}\widehat{a}_j Q_{t,j}\Big)^2,
\end{aligned}$$

where $\widehat{\boldsymbol{\Gamma}}_Q = (\widehat{\gamma}_Q(i,j) : 1 \le i,j \le p)$ is a $p \times p$ matrix and $\widehat{\boldsymbol{\gamma}}_Q = (\widehat{\gamma}_Q(0,1), \ldots, \widehat{\gamma}_Q(0,p))^\top$ is a vector in $\mathbb{R}^p$ with the entries $\widehat{\gamma}_Q(i,j) = \sum_{t=p+1}^{T} Q_{t,i}Q_{t,j}$.

The estimators $\widehat{\boldsymbol{a}}$ and $\widehat{\nu}$ have the following theoretical properties.

Proposition 4.2. *Suppose that the process $\{\eta_t\}$ has a finite fourth cumulant κ. Then*

$$\sqrt{T}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \xrightarrow{d} N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\nu^*\Gamma^{-1} & 0\\ 0 & 2(\nu^*)^2 + \kappa\end{pmatrix}\right),$$

where $\boldsymbol{\Gamma} = (\gamma(i-j) : 1 \le i,j \le p)$ is the autocovariance matrix of the AR(p) process $\{Y_t\}$ with $Y_t = \sum_{j=1}^{p} a_j^* Y_{t-j} + \eta_t$.

Derivation of (4.9) and (4.10). Writing $W_t(\boldsymbol{a}^*) = X_t - \sum_{j=1}^{p} a_j^* X_{t-j}$, we have

$$\boldsymbol{\xi}_t(\theta^*) = W_t(\boldsymbol{a}^*) + \Pi_{t-1}\eta_{t-1} \tag{4.11}$$

for $t > p$. By definition, $\Pi_t\eta_t$ belongs to the linear space spanned by $X_1, \ldots, X_{t-1}$. Moreover $\Pi_t\eta_t$ is orthogonal to the space spanned by $X_1, \ldots, X_{t-1}$ since $\Pi_{t-1}\Pi_t\eta_t = \Pi_{t-1}\eta_t = 0$. Noticing $\text{span}\{\xi_t(\boldsymbol{\theta}^*)\} \oplus \text{span}\{X_1, \ldots, X_{t-1}\} = \text{span}\{X_1, \ldots, X_t\}$, we can infer that

$$\Pi_t\eta_t = \frac{\mathbb{E}[\eta_t\xi_t(\boldsymbol{\theta}^*)]}{e_t(\boldsymbol{\theta}^*)}\xi_t(\boldsymbol{\theta}^*). \tag{4.12}$$

Since $\xi_t(\boldsymbol{\theta}^*) = \eta_t + (\Pi_{t-1}\eta_{t-1} - \eta_{t-1})$, it holds that $\mathbb{E}[\eta_t \xi_t(\boldsymbol{\theta}^*)] = \nu^*$ and $e_t(\boldsymbol{\theta}^*) = \nu^* + \mu_{t-1}$ with $\mu_t = \mathbb{E}[(\eta_t - \Pi_t \eta_t)^2]$. Plugging this into (4.12) yields

$$\xi_t(\boldsymbol{\theta}^*) = W_t(\boldsymbol{a}^*) + \frac{\nu^*}{\nu^* + \mu_{t-2}} \xi_{t-1}(\boldsymbol{\theta}^*). \tag{4.13}$$

The term μ_t can be rewritten as

$$\mu_t = \mathbb{E}[(\eta_t - \Pi_t \eta_t)^2] = \nu^* - \mathbb{E}[\Pi_t \eta_t] = \nu^* - \frac{(\nu^*)^2}{\nu^* + \mu_{t-1}} = \frac{\nu^* \mu_{t-1}}{\nu^* + \mu_{t-1}}.$$

This yields the recurrence equation $1/\mu_t = 1/\nu^* + 1/\mu_{t-1}$, which can be recursively applied to obtain that $1/\mu_t = (t-p)/\nu^* + 1/\mu_p$ for $t > p$. Using this in (4.13) gives that

$$\frac{\nu^*}{\nu^* + \mu_{t-2}} = \frac{\nu^*/\mu_{t-2}}{1 + \nu^*/\mu_{t-2}} = \frac{\nu^*/\mu_p + t - 2 - p}{\nu^*/\mu_p + t - 1 - p}$$

and thus

$$\xi_t(\boldsymbol{\theta}^*) = W_t(\boldsymbol{a}^*) + \frac{\beta(\nu^*) + t - 1}{\beta(\nu^*) + t} \xi_{t-1}(\boldsymbol{\theta}^*) \tag{4.14}$$

with $\beta(\nu^*) = \nu^*/\mu_p - p - 1$ for $t > p + 1$. By iteratively applying (4.14), we arrive at

$$\xi_t(\boldsymbol{\theta}^*) = \sum_{s=p+1}^{t} \frac{\beta(\nu^*) + s}{\beta(\nu^*) + t} W_s(\boldsymbol{a}^*) + \frac{\beta(\nu^*) + p + 1}{\beta(\nu^*) + t} \Pi_p \eta_p,$$

which can be equivalently written as

$$\xi_t(\boldsymbol{\theta}^*) = V_t(\boldsymbol{a}^*) - \frac{1}{\beta(\nu^*) + t} \sum_{s=p+1}^{t-1} V_s(\boldsymbol{a}^*) + \frac{\beta(\nu^*) + p + 1}{\beta(\nu^*) + t} \Pi_p \eta_p.$$

Moreover, using the representation $e_t(\boldsymbol{\theta}^*) = \nu^* + \mu_{t-1}$ and the formulas on μ_t from above, it is easily seen that

$$e_t(\boldsymbol{\theta}^*) = \Big(1 + \frac{1}{\beta(\nu^*) + t}\Big) \nu^*.$$

Proof of Proposition 4.2. Let the process $\{Y_t\}$ be defined by the equations $Y_t = \sum_{j=1}^{p} a_j^* Y_{t-j} + \eta_t$. Since $X_t = Y_t - Y_{t-1}$, we obtain that

$$V_t(\boldsymbol{a}) = \sum_{k=p+1}^{t} \Big(X_k - \sum_{j=1}^{p} a_j X_{k-j} \Big) \tag{4.15}$$

$$= \{Y_t - Y_p\} - \sum_{j=1}^{p} a_j \{Y_{t-j} - Y_{p-j}\}. \tag{4.16}$$

From (4.15), it immediately follows that

$$\widehat{\xi}_t(\boldsymbol{\theta}) = \eta_t(\boldsymbol{a}) - \frac{1}{t}\sum_{k=p+1}^{t-1}\eta_k(\boldsymbol{a}) - \frac{p+1}{t}\eta_p(\boldsymbol{a}) \quad \text{with} \quad \eta_t(\boldsymbol{a}) = Y_t - \sum_{j=1}^{p} a_j Y_{t-j}, \tag{4.17}$$

where $\eta_t(\boldsymbol{a})$ equals η_t for $\boldsymbol{a} = \boldsymbol{a}^*$, that is, $\eta_t(\boldsymbol{a}^*) = \eta_t$. With the help of (4.16), we can further write

$$\widehat{\xi}_t(\boldsymbol{\theta}) = U_{t,0} - \sum_{j=1}^{p} a_j U_{t,j} \quad \text{with} \quad U_{t,j} = Y_{t-j} - \frac{1}{t}\sum_{k=p+1}^{t-1} Y_{k-j} - \frac{p+1}{t} Y_{p-j}. \tag{4.18}$$

Using (4.18) and taking the first derivatives of the likelihood $L_t(\boldsymbol{\theta})$, we obtain the first-order conditions

$$\frac{\partial L_T(\boldsymbol{\theta})}{\partial a_k} = \frac{1}{\nu}\sum_{p+1}^{T}\Big(U_{t,0} - \sum_{j=1}^{p} a_j U_{t,j}\Big)U_{t,k} \stackrel{!}{=} 0 \quad \text{for } 1 \le k \le p \tag{4.19}$$

$$\frac{\partial L_T(\boldsymbol{\theta})}{\partial \nu} = -\frac{T-p}{2\nu} + \frac{1}{2\nu^2}\sum_{t=p+1}^{T}\widehat{\xi}_t^2(\boldsymbol{\theta}) \stackrel{!}{=} 0. \tag{4.20}$$

From (4.19) together with some straightforward calculations, we get that

$$\sum_{j=1}^{p}\Big(\frac{1}{T-p}\sum_{t=p+1}^{T} U_{t,j}U_{t,k}\Big)\big(\widehat{a}_j - a_j^*\big) = \frac{1}{T-p}\sum_{t=p+1}^{T}\widehat{\xi}_t(\theta^*)U_{t,k} \tag{4.21}$$

for $1 \le k \le p$, or equivalently,

$$\widehat{\boldsymbol{\Gamma}}_U(\widehat{a} - a^*) = \widehat{\boldsymbol{\rho}}_U, \tag{4.22}$$

where $\widehat{\boldsymbol{\rho}}_U = (\widehat{\rho}_U(1), \ldots, \widehat{\rho}_U(p))^\top$ with $\widehat{\rho}_U(k) = (T-p)^{-1}\sum_{t=p+1}^{T}\widehat{\xi}_t(\theta^*)U_{t,k}$ and

$$\widehat{\boldsymbol{\Gamma}}_U = \begin{pmatrix} \widehat{\gamma}_U(1,1) & \ldots & \widehat{\gamma}_U(p,1) \\ \vdots & \ddots & \vdots \\ \widehat{\gamma}_U(1,p) & \ldots & \widehat{\gamma}_U(p,p) \end{pmatrix}$$

with $\widehat{\gamma}_U(j,k) = (T-p)^{-1}\sum_{t=p+1}^{T} U_{t,j}U_{t,k}$. From (4.20) and (4.21), it further follows that

$$\widehat{\nu} = \frac{1}{T-p}\sum_{t=p+1}^{T}\widehat{\xi}_t^2(\boldsymbol{\theta}^*) - \sum_{j=1}^{p}(\widehat{a}_j - a_j^*)\Big(\frac{1}{T-p}\sum_{t=p+1}^{T}\widehat{\xi}_t(\boldsymbol{\theta}^*)U_{t,j}\Big). \tag{4.23}$$

Noting that $\partial\widehat{\xi}_t(\boldsymbol{\theta})/\partial\boldsymbol{\theta} = (-U_{t,1}, \ldots, -U_{t,p})$, Lemmas 5 and 6 in Pham-Dinh (1978)

yield that $\widehat{\mathbf{\Gamma}}_U = \mathbf{\Gamma} + o_p(1)$ and

$$\sqrt{T}\begin{pmatrix} \widehat{\boldsymbol{\rho}}_U \\ \frac{1}{T-p}\sum_{t=p+1}^{T}\widehat{\xi}_t^2(\boldsymbol{\theta}^*) \end{pmatrix} \xrightarrow{d} N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\nu^*\mathbf{\Gamma} & 0\\ 0 & 2(\nu^*)^2+\kappa\end{pmatrix}\right). \tag{4.24}$$

(To prove (4.24), one uses that $U_{t,j} = Y_{t-j} - t^{-1}\sum_{k=p+1}^{t-1} Y_{k-j} - \{(p+1)/t\}Y_{p-j}$ with $\{Y_t\}$ being a stationary, causal AR(p) process and $\widehat{\xi}_t(\boldsymbol{\theta}^*) = \eta_t - t^{-1}\sum_{k=p+1}^{t-1}\eta_k - \{(p+1)/t\}\eta_p$ with η_t being i.i.d. variables.) Proposition 4.2 follows upon applying these results to (4.22) and (4.23). □

Appendix

In what follows, we prove the theoretical results from Section 3. The proofs of the results from Section 4 are deferred to the Supplementary Material. Throughout the Appendix, we use the following notation: The symbol C denotes a universal real constant which may take a different value on each occurrence. For $a, b \in \mathbb{R}$, we write $a_+ = \max\{0, a\}$ and $a \vee b = \max\{a, b\}$. For any set A, the symbol $|A|$ denotes the cardinality of A. The notation $X \stackrel{\mathcal{D}}{=} Y$ means that the two random variables X and Y have the same distribution. Finally, $f_0(\cdot)$ and $F_0(\cdot)$ denote the density and distribution function of the standard normal distribution, respectively.

Auxiliary results using strong approximation theory

The main purpose of this section is to prove that there is a version of the multiscale statistic $\widehat{\Phi}_T$ defined in (3.4) which is close to a Gaussian statistic whose distribution is known. More specifically, we prove the following result.

Proposition A.1. *Under the conditions of Theorem 3.1, there exist statistics $\widetilde{\Phi}_T$ for $T = 1, 2, \ldots$ with the following two properties: (i) $\widetilde{\Phi}_T$ has the same distribution as $\widehat{\Phi}_T$ for any T, and (ii)*

$$\big|\widetilde{\Phi}_T - \Phi_T\big| = o_p\Big(\frac{T^{1/q}}{\sqrt{T h_{\min}}} + \rho_T \sqrt{\log T}\Big),$$

where Φ_T is a Gaussian statistic as defined in (3.3).

Proof of Proposition A.1. For the proof, we draw on strong approximation theory for stationary processes $\{\varepsilon_t\}$ that fulfill the conditions (C1)–(C3). By Theorem 2.1 and Corollary 2.1 in Berkes et al. (2014), the following strong approximation result holds true: On a richer probability space, there exist a standard Brownian motion $\mathbb{B}$ and a sequence $\{\widetilde{\varepsilon}_t : t \in \mathbb{N}\}$ such that $[\widetilde{\varepsilon}_1, \ldots, \widetilde{\varepsilon}_T] \stackrel{\mathcal{D}}{=} [\varepsilon_1, \ldots, \varepsilon_T]$ for each T and

$$\max_{1 \le t \le T} \Big| \sum_{s=1}^{t} \widetilde{\varepsilon}_s - \sigma \mathbb{B}(t) \Big| = o\big(T^{1/q}\big) \quad \text{a.s.}, \tag{A.1}$$

where $\sigma^2 = \sum_{k \in \mathbb{Z}} \mathrm{Cov}(\varepsilon_0, \varepsilon_k)$ denotes the long-run error variance. To apply this result, we define

$$\widetilde{\Phi}_T = \max_{(u,h) \in \mathcal{G}_T} \Big\{ \Big| \frac{\widetilde{\phi}_T(u,h)}{\widetilde{\sigma}} \Big| - \lambda(h) \Big\},$$

where $\widetilde{\phi}_T(u,h) = \sum_{t=1}^{T} w_{t,T}(u,h) \widetilde{\varepsilon}_t$ and $\widetilde{\sigma}^2$ is the same estimator as $\widehat{\sigma}^2$ with $Y_t = m(t/T) + \varepsilon_t$ replaced by $\widetilde{Y}_t = m(t/T) + \widetilde{\varepsilon}_t$ for $1 \le t \le T$. In addition, we let

$$\Phi_T = \max_{(u,h) \in \mathcal{G}_T} \Big\{ \Big| \frac{\phi_T(u,h)}{\sigma} \Big| - \lambda(h) \Big\}$$

$$\Phi_T^\diamond = \max_{(u,h)\in\mathcal{G}_T} \Big\{ \Big| \frac{\phi_T(u,h)}{\widetilde{\sigma}} \Big| - \lambda(h) \Big\}$$

with $\phi_T(u,h) = \sum_{t=1}^T w_{t,T}(u,h)\sigma Z_t$ and $Z_t = \mathbb{B}(t) - \mathbb{B}(t-1)$. With this notation, we can write

$$\big|\widetilde{\Phi}_T - \Phi_T\big| \le \big|\widetilde{\Phi}_T - \Phi_T^\diamond\big| + \big|\Phi_T^\diamond - \Phi_T\big| = \big|\widetilde{\Phi}_T - \Phi_T^\diamond\big| + o_p\big(\rho_T\sqrt{\log T}\big), \tag{A.2}$$

where the last equality follows by taking into account that $\phi_T(u,h) \sim N(0,\sigma^2)$ for all $(u,h) \in \mathcal{G}_T$, $|\mathcal{G}_T| = O(T^\theta)$ for some large but fixed constant θ and $\widetilde{\sigma}^2 = \sigma^2 + o_p(\rho_T)$. Straightforward calculations yield that

$$\big|\widetilde{\Phi}_T - \Phi_T^\diamond\big| \le \widetilde{\sigma}^{-1} \max_{(u,h)\in\mathcal{G}_T} \big|\widetilde{\phi}_T(u,h) - \phi_T(u,h)\big|.$$

Using summation by parts, we further obtain that

$$\begin{aligned}\big|\widetilde{\phi}_T(u,h) - \phi_T(u,h)\big| &\le W_T(u,h) \max_{1\le t\le T} \Big| \sum_{s=1}^t \widetilde{\varepsilon}_s - \sigma \sum_{s=1}^t \big\{\mathbb{B}(s) - \mathbb{B}(s-1)\big\}\Big| \\ &= W_T(u,h) \max_{1\le t\le T} \Big| \sum_{s=1}^t \widetilde{\varepsilon}_s - \sigma\mathbb{B}(t)\Big|,\end{aligned}$$

where

$$W_T(u,h) = \sum_{t=1}^{T-1} |w_{t+1,T}(u,h) - w_{t,T}(u,h)| + |w_{T,T}(u,h)|.$$

Standard arguments show that $\max_{(u,h)\in\mathcal{G}_T} W_T(u,h) = O(1/\sqrt{Th_{\min}})$. Applying the strong approximation result (A.1), we can thus infer that

$$\begin{aligned}\big|\widetilde{\Phi}_T - \Phi_T^\diamond\big| &\le \widetilde{\sigma}^{-1} \max_{(u,h)\in\mathcal{G}_T} \big|\widetilde{\phi}_T(u,h) - \phi_T(u,h)\big| \\ &\le \widetilde{\sigma}^{-1} \max_{(u,h)\in\mathcal{G}_T} W_T(u,h) \max_{1\le t\le T} \Big| \sum_{s=1}^t \widetilde{\varepsilon}_s - \sigma\mathbb{B}(t)\Big| = o_p\Big(\frac{T^{1/q}}{\sqrt{Th_{\min}}}\Big).\end{aligned} \tag{A.3}$$

Plugging (A.3) into (A.2) completes the proof. □

Auxiliary results using anti-concentration bounds

In this section, we establish some properties of the Gaussian statistic Φ_T defined in (3.3). We in particular show that Φ_T does not concentrate too strongly in small regions of the form $[x - \delta_T, x + \delta_T]$ with δ_T converging to zero.

Proposition A.2. *Under the conditions of Theorem 3.1, it holds that*

$$\sup_{x\in\mathbb{R}} \mathbb{P}\Big(|\Phi_T - x| \leq \delta_T\Big) = o(1),$$

where $\delta_T = T^{1/q}/\sqrt{Th_{\min}} + \rho_T\sqrt{\log T}$.

Proof of Proposition A.2. The main technical tool for proving Proposition A.2 are anti-concentration bounds for Gaussian random vectors. The following proposition slightly generalizes anti-concentration results derived in Chernozhukov et al. (2015), in particular Theorem 3 therein.

Proposition A.3. *Let* $(X_1,\dots,X_p)^\top$ *be a Gaussian random vector in* $\mathbb{R}^p$ *with* $\mathbb{E}[X_j] = \mu_j$ *and* $\mathrm{Var}(X_j) = \sigma_j^2 > 0$ *for* $1 \leq j \leq p$. *Define* $\overline{\mu} = \max_{1\leq j\leq p}|\mu_j|$ *together with* $\underline{\sigma} = \min_{1\leq j\leq p}\sigma_j$ *and* $\overline{\sigma} = \max_{1\leq j\leq p}\sigma_j$. *Moreover, set* $a_p = \mathbb{E}[\max_{1\leq j\leq p}(X_j-\mu_j)/\sigma_j]$ *and* $b_p = \mathbb{E}[\max_{1\leq j\leq p}(X_j-\mu_j)]$. *For every* $\delta > 0$, *it holds that*

$$\sup_{x\in\mathbb{R}} \mathbb{P}\Big(\big|\max_{1\leq j\leq p} X_j - x\big| \leq \delta\Big) \leq C\delta\big\{\overline{\mu} + a_p + b_p + \sqrt{1\vee\log(\underline{\sigma}/\delta)}\big\},$$

where $C > 0$ *depends only on* $\underline{\sigma}$ *and* $\overline{\sigma}$.

The proof of Proposition A.3 is provided in the Supplementary Material. To apply Proposition A.3 to our setting at hand, we introduce the following notation: We write $x = (u,h)$ along with $\mathcal{G}_T = \{x : x \in \mathcal{G}_T\} = \{x_1,\dots,x_p\}$, where $p := |\mathcal{G}_T| \leq O(T^\theta)$ for some large but fixed $\theta > 0$ by our assumptions. Moreover, for $j = 1,\dots,p$, we set

$$\begin{aligned} X_{2j-1} &= \frac{\phi_T(x_{j1},x_{j2})}{\sigma} - \lambda(x_{j2}) \\ X_{2j} &= -\frac{\phi_T(x_{j1},x_{j2})}{\sigma} - \lambda(x_{j2}) \end{aligned}$$

with $x_j = (x_{j1},x_{j2})$. This notation allows us to write

$$\Phi_T = \max_{1\leq j\leq 2p} X_j,$$

where $(X_1,\dots,X_{2p})^\top$ is a Gaussian random vector with the following properties: (i) $\mu_j := \mathbb{E}[X_j] = -\lambda(x_{j2})$ and thus $\overline{\mu} = \max_{1\leq j\leq 2p}|\mu_j| \leq C\sqrt{\log T}$, and (ii) $\sigma_j^2 := \mathrm{Var}(X_j) = 1$ for all j. Since $\sigma_j = 1$ for all j, it holds that $a_{2p} = b_{2p}$. Moreover, as the variables $(X_j-\mu_j)/\sigma_j$ are standard normal, we have that $a_{2p} = b_{2p} \leq \sqrt{2\log(2p)} \leq C\sqrt{\log T}$. With this notation at hand, we can apply Proposition A.3 to obtain that

$$\sup_{x\in\mathbb{R}} \mathbb{P}\Big(\big|\Phi_T - x\big| \leq \delta_T\Big) \leq C\delta_T\Big[\sqrt{\log T} + \sqrt{\log(1/\delta_T)}\Big] = o(1)$$

with $\delta_T = T^{1/q}/\sqrt{Th_{\min}} + \rho_T\sqrt{\log T}$, which is the statement of Proposition A.2. □

Proof of Theorem 3.1

To prove Theorem 3.1, we make use of the two auxiliary results derived above. By Proposition A.1, there exist statistics $\widetilde{\Phi}_T$ for $T = 1, 2, \ldots$ which are distributed as $\widehat{\Phi}_T$ for any $T \geq 1$ and which have the property that

$$\big|\widetilde{\Phi}_T - \Phi_T\big| = o_p\Big(\frac{T^{1/q}}{\sqrt{T h_{\min}}} + \rho_T \sqrt{\log T}\Big), \tag{A.4}$$

where Φ_T is a Gaussian statistic as defined in (3.3). The approximation result (A.4) allows us to replace the multiscale statistic $\widehat{\Phi}_T$ by an identically distributed version $\widetilde{\Phi}_T$ which is close to the Gaussian statistic Φ_T. In the next step, we show that

$$\sup_{x \in \mathbb{R}} \big|\mathbb{P}(\widetilde{\Phi}_T \leq x) - \mathbb{P}(\Phi_T \leq x)\big| = o(1), \tag{A.5}$$

which immediately implies the statement of Theorem 3.1. For the proof of (A.5), we use the following simple lemma:

Lemma A.4. *Let V_T and W_T be real-valued random variables for $T = 1, 2, \ldots$ such that $V_T - W_T = o_p(\delta_T)$ with some $\delta_T = o(1)$. If*

$$\sup_{x \in \mathbb{R}} \mathbb{P}(|V_T - x| \leq \delta_T) = o(1), \tag{A.6}$$

then

$$\sup_{x \in \mathbb{R}} \big|\mathbb{P}(V_T \leq x) - \mathbb{P}(W_T \leq x)\big| = o(1). \tag{A.7}$$

The statement of Lemma A.4 can be summarized as follows: If W_T can be approximated by V_T in the sense that $V_T - W_T = o_p(\delta_T)$ and if V_T does not concentrate too strongly in small regions of the form $[x - \delta_T, x + \delta_T]$ as assumed in (A.6), then the distribution of W_T can be approximated by that of V_T in the sense of (A.7).

Proof of Lemma A.4. It holds that

$$\begin{aligned}
&\big|\mathbb{P}(V_T \leq x) - \mathbb{P}(W_T \leq x)\big| \\
&= \big|\mathbb{E}\big[1(V_T \leq x) - 1(W_T \leq x)\big]\big| \\
&\leq \big|\mathbb{E}\big[\big\{1(V_T \leq x) - 1(W_T \leq x)\big\} 1(|V_T - W_T| \leq \delta_T)\big]\big| + \big|\mathbb{E}\big[1(|V_T - W_T| > \delta_T)\big]\big| \\
&\leq \mathbb{E}\big[1(|V_T - x| \leq \delta_T, |V_T - W_T| \leq \delta_T)\big] + o(1) \\
&\leq \mathbb{P}(|V_T - x| \leq \delta_T) + o(1).
\end{aligned}$$

□

We now apply this lemma with $V_T = \Phi_T$, $W_T = \widetilde{\Phi}_T$ and $\delta_T = T^{1/q}/\sqrt{T h_{\min}} + \rho_T \sqrt{\log T}$: From (A.4), we already know that $\widetilde{\Phi}_T - \Phi_T = o_p(\delta_T)$. Moreover, by Proposition A.2, it holds that

$$\sup_{x \in \mathbb{R}} \mathbb{P}\Big(|\Phi_T - x| \leq \delta_T\Big) = o(1). \tag{A.8}$$

Hence, the conditions of Lemma A.4 are satisfied. Applying the lemma, we obtain (A.5), which completes the proof of Theorem 3.1.

Proof of Proposition 3.2

To start with, we introduce the notation $\widehat{\psi}_T(u,h) = \widehat{\psi}_T^A(u,h) + \widehat{\psi}_T^B(u,h)$, where $\widehat{\psi}_T^A(u,h) = \sum_{t=1}^T w_{t,T}(u,h)\varepsilon_t$ and $\widehat{\psi}_T^B(u,h) = \sum_{t=1}^T w_{t,T}(u,h)m_T(\frac{t}{T})$. We further write $m_T(\frac{t}{T}) = m_T(u) + m_T'(\xi_{u,t,T})(\frac{t}{T} - u)$, where $\xi_{u,t,T}$ is an intermediate point between u and t/T. By assumption, there exists $(u_0, h_0) \in \mathcal{G}_T$ with $[u_0 - h_0, u_0 + h_0] \subseteq [0,1]$ such that $m_T'(w) \geq c_T\sqrt{\log T/(Th_0^3)}$ for all $w \in [u_0 - h_0, u_0 + h_0]$. (The case that $-m_T'(w) \geq c_T\sqrt{\log T/(Th_0^3)}$ for all w can be treated analogously.) Below, we prove that under this assumption,

$$\widehat{\psi}_T^B(u_0, h_0) \geq \frac{\kappa c_T\sqrt{\log T}}{2}, \tag{A.9}$$

where $\kappa = (\int K(\varphi)\varphi^2 d\varphi)/(\int K^2(\varphi)\varphi^2 d\varphi)^{1/2}$. Moreover, by arguments very similar to those for the proof of Proposition A.1, it follows that

$$\max_{(u,h)\in\mathcal{G}_T} |\widehat{\psi}_T^A(u,h)| = O_p(\sqrt{\log T}). \tag{A.10}$$

With the help of (A.9), (A.10) and the fact that $\lambda(h) \leq \lambda(h_{\min}) \leq C\sqrt{\log T}$, we can infer that

$$\begin{aligned}
\widehat{\Psi}_T &\geq \max_{(u,h)\in\mathcal{G}_T} \frac{|\widehat{\psi}_T^B(u,h)|}{\widehat{\sigma}} - \max_{(u,h)\in\mathcal{G}_T} \Big\{\frac{|\widehat{\psi}_T^A(u,h)|}{\widehat{\sigma}} + \lambda(h)\Big\} \\
&= \max_{(u,h)\in\mathcal{G}_T} \frac{|\widehat{\psi}_T^B(u,h)|}{\widehat{\sigma}} + O_p(\sqrt{\log T}) \\
&\geq \frac{\kappa c_T\sqrt{\log T}}{2\widehat{\sigma}} + O_p(\sqrt{\log T}).
\end{aligned} \tag{A.11}$$

Since $q_T(\alpha) = O(\sqrt{\log T})$ for any fixed $\alpha \in (0,1)$, (A.11) immediately yields that $\mathbb{P}(\widehat{\Psi}_T \leq q_T(\alpha)) = o(1)$, which is the statement of Proposition 3.2.

Proof of (A.9). Since the kernel K is symmetric and $u_0 = t/T$ for some t, it holds that $S_{T,1}(u_0, h_0) = 0$, which in turn implies that

$$\begin{aligned}
&w_{t,T}(u_0,h_0)\Big(\frac{\frac{t}{T} - u_0}{h_0}\Big) \\
&\quad = K\Big(\frac{\frac{t}{T} - u_0}{h_0}\Big)\Big(\frac{\frac{t}{T} - u_0}{h_0}\Big)^2 \Big/ \Big\{\sum_{t=1}^T K^2\Big(\frac{\frac{t}{T} - u_0}{h_0}\Big)\Big(\frac{\frac{t}{T} - u_0}{h_0}\Big)^2\Big\}^{1/2} \geq 0.
\end{aligned}$$

From this and the assumption that $m_T'(w) \geq c_T\sqrt{\log T/(Th_0^3)}$ for all $w \in [u_0 - h_0, u_0 +$

$h_0]$, we get that

$$\widehat{\psi}_T^B(u_0,h_0) \geq c_T\sqrt{\frac{\log T}{Th_0}}\sum_{t=1}^{T} w_{t,T}(u_0,h_0)\Big(\frac{\frac{t}{T}-u_0}{h_0}\Big). \tag{A.12}$$

Standard calculations exploiting the Lipschitz continuity of the kernel K show that for any $(u,h) \in \mathcal{G}_T$ and any given natural number ℓ,

$$\Big|\frac{1}{Th}\sum_{t=1}^{T} K\Big(\frac{\frac{t}{T}-u}{h}\Big)\Big(\frac{\frac{t}{T}-u}{h}\Big)^{\ell} - \int_0^1 \frac{1}{h}K\Big(\frac{w-u}{h}\Big)\Big(\frac{w-u}{h}\Big)^{\ell} dw\Big| \leq \frac{C}{Th}, \tag{A.13}$$

where the constant C does not depend on u, h and T. With the help of (A.13), we obtain that for any $(u,h) \in \mathcal{G}_T$ with $[u-h,u+h] \subseteq [0,1]$,

$$\Big|\sum_{t=1}^{T} w_{t,T}(u,h)\Big(\frac{\frac{t}{T}-u}{h}\Big) - \frac{\sqrt{Th}}{\kappa}\Big| \leq \frac{C}{\sqrt{Th}}, \tag{A.14}$$

where the constant C does again not depend on u, h and T. (A.14) implies that $\sum_{t=1}^{T} w_{t,T}(u,h)(\frac{t}{T}-u)/h \geq \kappa\sqrt{Th}/2$ for sufficiently large T and any $(u,h) \in \mathcal{G}_T$ with $[u-h,u+h] \subseteq [0,1]$. Using this together with (A.12), we immediately obtain (A.9). $\square$

Proof of Proposition 3.3

In what follows, we show that

$$\mathbb{P}(E_T^+) \geq (1-\alpha) + o(1). \tag{A.15}$$

The other statements of Proposition 3.3 can be verified by analogous arguments. (A.15) is a consequence of the following two observations:

(i) For all $(u,h) \in \mathcal{G}_T$ with

$$\Big|\frac{\widehat{\psi}_T(u,h) - \mathbb{E}\widehat{\psi}_T(u,h)}{\widehat{\sigma}}\Big| - \lambda(h) \leq q_T(\alpha) \quad \text{and} \quad \frac{\widehat{\psi}_T(u,h)}{\widehat{\sigma}} - \lambda(h) > q_T(\alpha),$$

it holds that $\mathbb{E}[\widehat{\psi}_T(u,h)] > 0$.

(ii) For all $(u,h) \in \mathcal{G}_T$ with $[u-h,u+h] \subseteq [0,1]$, $\mathbb{E}[\widehat{\psi}_T(u,h)] > 0$ implies that $m'(v) > 0$ for some $v \in [u-h,u+h]$.

Observation (i) is trivial, (ii) can be seen as follows: Let (u,h) be any point with $(u,h) \in \mathcal{G}_T$ and $[u-h,u+h] \subseteq [0,1]$. It holds that $\mathbb{E}[\widehat{\psi}_T(u,h)] = \widehat{\psi}_T^B(u,h)$, where $\widehat{\psi}_T^B(u,h)$ has been defined in the proof of Proposition 3.2. There, we have already seen

that

$$\widehat{\psi}_T^B(u,h) = \sum_{t=1}^{T} w_{t,T}(u,h)\Big(\frac{\frac{t}{T}-u}{h}\Big)\, hm'(\xi_{u,t,T}),$$

where $\xi_{u,t,T}$ is some intermediate point between u and t/T. Moreover, $S_{T,1}(u,h) = 0$, which implies that $w_{t,T}(u,h)(\frac{t}{T}-u)/h \geq 0$ for any t. Hence, $\mathbb{E}[\widehat{\psi}_T(u,h)] = \widehat{\psi}_T^B(u,h)$ can only take a positive value if $m'(v) > 0$ for some $v \in [u-h, u+h]$.
From observations (i) and (ii), we can draw the following conclusions: On the event

$$\big\{\widehat{\Phi}_T \leq q_T(\alpha)\big\} = \Big\{ \max_{(u,h)\in\mathcal{G}_T} \Big(\Big|\frac{\widehat{\psi}_T(u,h) - \mathbb{E}\widehat{\psi}_T(u,h)}{\widehat{\sigma}}\Big| - \lambda(h)\Big) \leq q_T(\alpha)\Big\},$$

it holds that for all $(u,h) \in \mathcal{A}_T^+$, $m'(v) > 0$ for some $v \in I_{u,h} = [u-h, u+h]$. We thus obtain that $\{\widehat{\Phi}_T \leq q_T(\alpha)\} \subseteq E_T^+$. This in turn implies that

$$\mathbb{P}(E_T^+) \geq \mathbb{P}\big(\widehat{\Phi}_T \leq q_T(\alpha)\big) = (1-\alpha) + o(1),$$

where the last equality holds by Theorem 3.1.

References

Andrews, D. W. K. (1991). Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica*, **59** 817–858.

Benner, T. C. (1999). Central england temperatures: long-term variability and teleconnections. *International Journal of Climatology*, **19** 391–403.

Berkes, I., Liu, W. and Wu, W. B. (2014). Komlós-Major-Tusnády approximation under dependence. *Annals of Probability*, **42** 794–817.

Chaudhuri, P. and Marron, J. S. (1999). SiZer for the exploration of structures in curves. *Journal of the American Statistical Association*, **94** 807–823.

Chaudhuri, P. and Marron, J. S. (2000). Scale space view of curve estimation. *Annals of Statistics*, **28** 408–428.

Chernozhukov, V., Chetverikov, D. and Kato, K. (2014). Gaussian approximation of suprema of empirical processes. *Annals of Statistics*, **42** 1564–1597.

Chernozhukov, V., Chetverikov, D. and Kato, K. (2015). Comparison and anti-concentration bounds for maxima of Gaussian random vectors. *Probability Theory and Related Fields*, **162** 47–70.

Chernozhukov, V., Chetverikov, D. and Kato, K. (2017). Central limit theorems and bootstrap in high dimensions. *Annals of Probability*, **45** 2309–2352.

Cho, H. and Fryzlewicz, P. (2012). Multiscale and multilevel technique for consistent segmentation of nonstationary time series. *Statistica Sinica*, **22** 207–229.

De Jong, R. M. and Davidson, J. (2000). Consistency of kernel estimators of heteroscedastic and autocorrelated covariance matrices. *Econometrica*, **68** 407–423.

Donoho, D., Johnstone, I., Kerkyacharian, G. and Picard, D. (1995). Wavelet shrinkage: Asymptopia? *Journal of the Royal Statistical Society: Series B*, **57** 301–369.

Dümbgen, L. (2002). Application of local rank tests to nonparametric regression. *Journal of Nonparametric Statistics*, **14** 511–537.

Dümbgen, L. and Spokoiny, V. G. (2001). Multiscale testing of qualitative hypotheses. *Annals of Statistics*, **29** 124–152.

Dümbgen, L. and Walther, G. (2008). Multiscale inference about a density. *Annals of Statistics*, **36** 1758–1785.

Eckle, K., Bissantz, N. and Dette, H. (2017). Multiscale inference for multivariate deconvolution. *Electronic Journal of Statistics*, **11** 4179–4219.

Hall, P. and Heckman, N. E. (2000). Testing for monotonicity of a regression mean by calibrating for linear functions. *Annals of Statistics*, **28** 20–39.

Hall, P. and Van Keilegom, I. (2003). Using difference-based methods for inference in nonparametric regression with time series errors. *Journal of the Royal Statistical Society: Series B*, **65** 443–456.

Hannig, J. and Marron, J. S. (2006). Advanced distribution theory for SiZer. *Journal of the American Statistical Association*, **101** 484–499.

Hart, J. D. (1989). Differencing as an approximate de-trending device. *Stochastic Processes and their Applications*, **31** 251–259.

Hart, J. D. (1991). Kernel regression estimation with time series errors. *Journal of the Royal Statistical Society: Series B*, **53** 173–187.

Hart, J. D. (1994). Automated kernel smoothing of dependent data by using time series cross-validation. *Journal of the Royal Statistical Society: Series B*, **56** 529–542.

Herrmann, E., Gasser, T. and Kneip, A. (1992). Choice of bandwidth for kernel regression when residuals are correlated. *Biometrika*, **79** 783–795.

Liu, W. and Wu, W. B. (2010). Asymptotics of spectral density estimates. *Econometric Theory*, **26** 1218–1245.

Müller, H.-G. and Stadtmüller, U. (1988). Detecting dependencies in smooth regression models. *Biometrika*, **75** 639–650.

Park, C., Marron, J. S. and Rondonotti, V. (2004). Dependent SiZer: goodness-of-fit tests for time series models. *Journal of Applied Statistics*, **31** 999–1017.

Pham-Dinh, T. (1978). Estimation of parameters in the ARMA model when the characteristic polynomial of the MA operator has a unit zero. *Annals of Statistics*, **6** 1369–1389.

Proksch, K., Werner, F. and Munk, A. (2018). Multiscale scanning in inverse problems.

Forthcoming in Annals of Statistics.

Qiu, D., Shao, Q. and Yang, L. (2013). Efficient inference for autoregressive coefficients in the presence of trends. *Journal of Multivariate Analysis*, **114** 40–53.

Rahmstorf, S., Foster, G. and Cahill, N. (2017). Global temperature evolution: recent trends and some pitfalls. *Environmental Research Letters*, **12**.

Rohde, A. (2008). Adaptive goodness-of-fit tests based on signed ranks. *Annals of Statistics*, **36** 1346–1374.

Rondonotti, V., Marron, J. S. and Park, C. (2007). SiZer for time series: a new approach to the analysis of trends. *Electronic Journal of Statistics*, **1** 268–289.

Rufibach, K. and Walther, G. (2010). The block criterion for multiscale inference about a density, with applications to other multiscale problems. *Journal of Computational and Graphical Statistics*, **19** 175–190.

Schmidt-Hieber, J., Munk, A. and Dümbgen, L. (2013). Multiscale methods for shape constraints in deconvolution: confidence statements for qualitative features. *Annals of Statistics*, **41** 1299–1328.

Shao, Q. and Yang, L. J. (2011). Autoregressive coefficient estimation in nonparametric analysis. *Journal of Time Series Analysis*, **32** 587–597.

Tecuapetla-Gómez, I. and Munk, A. (2017). Autocovariance estimation in regression with a discontinuous signal and m-dependent errors: a difference-based approach. *Scandinavian Journal of Statistics*, **44** 346–368.

Truong, Y. K. (1991). Nonparametric curve estimation with time series errors. *Journal of Statistical Planning and Inference*, **28** 167–183.

Von Sachs, R. and MacGibbon, B. (2000). Non-parametric curve estimation by Wavelet thresholding with locally stationary errors. *Scandinavian Journal of Statistics*, **27** 475–499.

Wu, W. B. (2005). Nonlinear system theory: another look at dependence. *Proc. Natn. Acad. Sci. USA*, **102** 14150–14154.

Wu, W. B. and Shao, X. (2004). Limit theorems for iterated random functions. *Journal of Applied Probability* 425–436.