Multiscale Inference for Nonparametric Time Trends

Marina Khismatullina
University of Bonn

Michael Vogt
University of Bonn

April 2, 2018

1 The model

The model setting for the test problems considered in Sections 2 and 3 is as follows: We observe a time series $\{Y_t : 1 \leq t \leq T\}$ of length T which satisfies the model equation

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

for $1 \leq t \leq T$. Here, m is an unknown nonparametric regression 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 $G : \mathbb{R}^{\mathbb{Z}} \to \mathbb{R}$ 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

$$\tau_q = \frac{q^2 - 4 + (q-2)\sqrt{q^2 + 20q + 4}}{8q}$$

and $A > \frac{2}{3}(1/q + 1 + \tau_q)$.

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.

The model setting for the test problem analyzed in Section 4 is closely related to the setting discussed above. The main difference is that we observe several rather than only one time series. In particular, we observe time series $\mathcal{Y}_i = \{Y_{it} : 1 \leq t \leq T\}$ of length T for $1 \leq i \leq n$. Each time series $\mathcal{Y}_i$ satisfies the regression equation

$$Y_{it} = m_i\Big(\frac{t}{T}\Big) + \alpha_i + \varepsilon_{it} \tag{1.2}$$

for $1 \leq t \leq T$, where m_i is an unknown nonparametric function defined on $[0,1]$, α_i is a (deterministic or random) intercept term and $\mathcal{E}_i = \{\varepsilon_{it} : 1 \leq t \leq T\}$ is a zero-mean stationary error process. For identification, we normalize the functions m_i such that $\int_0^1 m_i(u)du = 0$ for all $1 \leq i \leq n$. The term α_i can also be regarded as an additional error component. In the econometrics literature, it is commonly called a fixed effect error term. It can be interpreted as capturing unobserved characteristics of the time series $\mathcal{Y}_i$ which remain constant over time. We allow the error terms α_i to be dependent across i in an arbitrary way. Hence, by including them in model equation (1.2), we allow the n time series $\mathcal{Y}_i$ in our sample to be correlated with each other. More specifically, since $\mathrm{Cov}(Y_{it}, Y_{jt}) = \mathrm{Cov}(\alpha_i, \alpha_j)$, we can accommodate any correlations across i that do not change over time. Whereas the terms α_i may be correlated, the error processes $\mathcal{E}_i$ are assumed to be independent across i. In addition, each process $\mathcal{E}_i$ is supposed to satisfy the conditions (C1)–(C3). Finally note that throughout the paper, we assume the number of time series n in model (1.2) to be fixed. It is however possible to extend our theoretical results to the case where n slowly grows with the sample size T.

2 The multiscale method

In this section, we introduce our multiscale test method and the underlying theory for the simple hypothesis $H_0 : m = 0$ in model (1.1). Both the method and the theory for this simple case can be easily adapted to more interesting test problems as we will see in Sections 3 and 4.

2.1 Construction of the test statistic

To construct a multiscale test statistic for the hypothesis $H_0 : m = 0$ in model (1.1), we consider the kernel averages

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

where $w_{t,T}(u,h)$ is a kernel weight with $u \in [0,1]$ and the bandwidth parameter h. In order to avoid boundary issues, we work with a local linear weighting scheme. We in particular set

$$w_{t,T}(u,h) = \frac{\Lambda_{t,T}(u,h)}{\{\sum_{t=1}^{T} \Lambda_{t,T}^2(u,h)\}^{1/2}}, \tag{2.1}$$

where

$$\Lambda_{t,T}(u,h) = K\Big(\frac{\frac{t}{T}-u}{h}\Big)\Big[S_{T,2}(u,h) - S_{T,1}(u,h)\Big(\frac{\frac{t}{T}-u}{h}\Big)\Big],$$

$S_{T,\ell}(u,h) = (Th)^{-1}\sum_{t=1}^{T} K(\frac{\frac{t}{T}-u}{h})(\frac{\frac{t}{T}-u}{h})^{\ell}$ 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, $|K(v) - K(w)| \le C|v-w|$ for any $v, w \in \mathbb{R}$ and some constant $C > 0$.

Alternatively to the local linear weights defined in (2.1), we could also work with local constant weights which are defined analogously with $\Lambda_{t,T}(u,h) = K(\frac{\frac{t}{T}-u}{h})$. We however prefer to use local linear weights as these have superior theoretical properties at the boundary.

The kernel average $\widehat{\psi}_T(u,h)$ is a local average of the observations $Y_1, \ldots, Y_T$ which gives positive weight only to data points Y_t with $t/T \in [u-h, u+h]$. Hence, only observations Y_t with t/T close to the location u are taken into account, the amount of localization being determined by the bandwidth h. With the weights defined in (2.1), the kernel average $\widehat{\psi}_T(u,h)$ is nothing else than a rescaled local linear estimator of $m(u)$ with bandwidth h. The weights are chosen such that in the case of independent error terms ε_t, $\mathrm{Var}(\widehat{\psi}_T(u,h)) = \sigma^2$ for any location u and bandwidth h, where $\sigma^2 = \mathrm{Var}(\varepsilon_t)$. In the more general case that the error terms satisfy the weak dependence conditions

from Section 1, 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 $Th \to \infty$, where $\sigma^2 = \sum_{\ell=-\infty}^{\infty} \mathrm{Cov}(\varepsilon_0, \varepsilon_\ell)$ is the long-run variance of the error terms. Hence, the statistics $\widehat{\psi}_T(u,h)$ have approximately the same variance across u and h for sufficiently large sample sizes T. In what follows, we consider normalized versions $\widehat{\psi}_T(u,h)/\widehat{\sigma}$ of the kernel averages $\widehat{\psi}_T(u,h)$, where $\widehat{\sigma}^2$ is an estimator of the long-run error variance σ^2. The problem of estimating σ^2 is discussed in detail in Section 5. There, we construct estimators $\widehat{\sigma}^2$ with the property that $\widehat{\sigma}^2 = \sigma^2 + O_p(T^{-1/2})$ under appropriate conditions. For the time being, we suppose that $\widehat{\sigma}^2$ is an estimator with reasonable theoretical properties. We in particular assume that $\widehat{\sigma}^2 = \sigma^2 + o_p(\rho_T)$, where the convergence rate ρ_T is specified in Theorem 2.1 below and may be much slower than $T^{-1/2}$ for our theory to work.

Our multiscale statistic combines the kernel averages $\widehat{\psi}_T(u,h)$ for a wide range of different locations u and bandwidths or scales h. Specifically, it is defined as

$$\widehat{\Psi}_T = \max_{(u,h)\in\mathcal{G}_T} \Big\{ \Big| \frac{\widehat{\psi}_T(u,h)}{\widehat{\sigma}} \Big| - \lambda(h) \Big\}, \tag{2.2}$$

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 discussed 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 follow the approach pioneered by Dümbgen and Spokoiny (2001) and subtract the additive correction term $\lambda(h)$ from the statistics $\widehat{\psi}_T(u,h)/\widehat{\sigma}$ that correspond to the bandwidth level h. 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 null hypothesis $H_0 : m = 0$ holds true. 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 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 $\max_k \widehat{\psi}_T(x_k,h_\ell)/\sigma$ is approximately of size $\lambda(h_\ell)$ for small bandwidths h_ℓ. As $\lambda(h) \to \infty$ for $h \to 0$, this implies that $\max_k \widehat{\psi}_T(x_k,h_\ell)/\sigma$ tends to be much

larger in size for small than for large bandwidth values. As a result, the stochastic behaviour of the uncorrected statistic $\widehat{\Psi}_{T,\text{uncorrected}}$ tends to be dominated by the statistics $\widehat{\psi}_T(x_k, h_\ell)$ corresponding to small bandwidths h_ℓ. The additively corrected statistic $\widehat{\Psi}_T$, in contrast, puts the statistics $\widehat{\psi}_T(x_k, h_\ell)$ corresponding to different bandwidth values 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 $\mathcal{G} = \{(u, h) : u \in [0, 1] \text{ and } h \in [h_{\min}, h_{\max}]\}$, 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 $h_{\max} < 1/2$.

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

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

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. In particular, we can choose the minimal bandwidth $h_{\min}$ to be of the order $T^{-1/2}$ for any $q > 4$, which means that we can let $h_{\min}$ converge to 0 very quickly. Moreover, the maximal bandwidth $h_{\max}$ need not even converge to 0, which implies that we can pick it very large.

2.2 The test procedure

In order to formulate a test for the hypothesis $H_0 : m = 0$, we still need to specify a critical value. To do so, we define the statistic

$$\Phi_T = \max_{(u,h)\in\mathcal{G}_T} \Big\{ \Big| \frac{\phi_T(u, h)}{\sigma} \Big| - \lambda(h) \Big\}, \tag{2.3}$$

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 $H_0 : m = 0$ is now defined as follows: For a given significance level $\alpha \in (0,1)$, we reject H_0 if $\widehat{\Psi}_T > q_T(\alpha)$.

2.3 Theoretical properties of the test

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

$$\begin{aligned}\widehat{\Phi}_T &= \max_{(u,h)\in\mathcal{G}_T} \Big\{ \Big| \frac{\widehat{\psi}_T(u,h) - \mathbb{E}\widehat{\psi}_T(u,h)}{\widehat{\sigma}} \Big| - \lambda(h) \Big\} \\ &= \max_{(u,h)\in\mathcal{G}_T} \Big\{ \Big| \frac{\widehat{\phi}_T(u,h)}{\widehat{\sigma}} \Big| - \lambda(h) \Big\} \end{aligned} \tag{2.4}$$

with

$$\widehat{\phi}_T(u,h) = \sum_{t=1}^{T} w_{t,T}(u,h)\varepsilon_t.$$

According to the following theorem, the (known) quantile $q_T(\alpha)$ of Φ_T defined in Section 2.2 can be used as a proxy for the $(1-\alpha)$-quantile of the statistic $\widehat{\Phi}_T$.

Theorem 2.1. *Let (C1)–(C6) be fulfilled and assume that* $\widehat{\sigma}^2 = \sigma^2 + o_p(\rho_T)$ *with* $\rho_T = T^{1/q}/\sqrt{T h_{\min} \log T}$. *Then*

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

A full proof of Theorem 2.1 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

$$\big|\widetilde{\Phi}_T - \Phi_T\big| = o_p\Big(\frac{T^{1/q}}{\sqrt{T h_{\min}}}\Big), \tag{2.5}$$

where the Gaussian statistic Φ_T is defined in Section 2.2. 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

$$\sup_{x\in\mathbb{R}} \big|\mathbb{P}(\widetilde{\Phi}_T \le x) - \mathbb{P}(\Phi_T \le x)\big| = o(1), \tag{2.6}$$

which implies that for any given $\alpha \in (0,1)$, the known quantile $q_T(\alpha)$ of the Gaussian statistic Φ_T can be used as a proxy for the $(1-\alpha)$-quantile of the statistic $\widetilde{\Phi}_T$. The main tool for verifying (2.6) are anti-concentration results for Gaussian random vectors as derived in Chernozhukov et al. (2015). Combining (2.5) and (2.6), we arrive at the statement of Theorem 2.1.

With the help of Theorem 2.1, we can investigate the theoretical properties of our multiscale test. The first result is an immediate consequence of Theorem 2.1. It says that the test has the correct (asymptotic) size.

Proposition 2.2. *Let the conditions of Theorem 2.1 be satisfied. Under the null hypothesis $H_0 : m = 0$, it holds that*

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

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

$$m_T(w) \ge c_T \sqrt{\frac{\log T}{Th}} \quad \text{for all } w \in [u-h, u+h], \tag{2.7}$$

where $\{c_T\}$ is any sequence of positive numbers with $c_T \to \infty$. Alternatively to (2.7), we may also assume that $-m_T(w) \ge c_T\sqrt{\log T/(Th)}$ 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 (2.7).

Proposition 2.3. *Let the conditions of Theorem 2.1 be satisfied and consider any sequence of functions m_T with the property* (2.7)*. Then*

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

The proof of Proposition 2.3 can be found in the Appendix. To formulate the next result, we define

$$\Pi_T = \big\{I_{u,h} = [u-h, u+h] : (u,h) \in \mathcal{A}_T\big\}$$

with

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

Π_T 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 $q_T(\alpha)$. With this notation at

hand, we consider the event

$$E_T = \Big\{\forall I_{u,h} \in \Pi_T : m(v) \neq 0 \text{ for some } v \in I_{u,h} = [u-h, u+h]\Big\}.$$

This is the event that the null hypothesis is violated on all intervals $I_{u,h}$ for which the (corrected) test statistic $|\widehat{\psi}_T(u,h)/\widehat{\sigma}| - \lambda(h)$ is above the critical value $q_T(\alpha)$. We can make the following formal statement about the event E_T whose proof is given in the Appendix.

Proposition 2.4. *Under the conditions of Theorem 2.1, it holds that*

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

According to Proposition 2.4, our test procedure allows us to make uniform confidence statements of the following form: With (asymptotic) probability $\geq (1-\alpha)$, the null hypothesis $H_0 : m = 0$ is violated on all intervals $I_{u,h} \in \Pi_T$. Hence, our multiscale test does not only allow us to check whether the null hypothesis is violated. It also allows us to identify the regions where violations occur with a pre-specified level of confidence.

The statement of Proposition 2.4 suggests to graphically present the results of our multiscale test by plotting the intervals $I_{u,h} \in \Pi_T$, that is, by plotting the intervals where (with asymptotic probability $\geq 1-\alpha$) our test detects a violation of the null hypothesis. The drawback of this graphical presentation is that the number of intervals in Π_T is often quite large. To obtain a better graphical summary of the results, we replace Π_T by a subset $\Pi_T^{\min}$ which is constructed as follows: As in Dümbgen (2002), we call an interval $I_{u,h} \in \Pi_T$ minimal if there is no other interval $I_{u',h'} \in \Pi_T$ with $I_{u',h'} \subset I_{u,h}$. Let $\Pi_T^{\min}$ be the collection of all minimal intervals in Π_T and define the event

$$E_T^{\min} = \Big\{\forall I_{u,h} \in \Pi_T^{\min} : m(v) \neq 0 \text{ for some } v \in I_{u,h} = [u-h, u+h]\Big\}.$$

It is easily seen that $E_T = E_T^{\min}$. Hence, by Proposition 2.4, it holds that

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

This suggests to plot the minimal intervals in $\Pi_T^{\min}$ rather than the whole collection of intervals Π_T as a graphical summary of the test results. We in particular use this way of presenting the test results in our application examples of Section ??.

3 Testing for shape constraints of a time trend

In what follows, we construct a multiscale test for the null hypothesis that the trend function m in model (1.1) is constant. To achieve this, we adapt the methodology developed in Section 2. Importantly, the resulting multiscale procedure does not only allow to test whether the null hypothesis is violated. As we will see, it also allows to identify, with a certain statistical confidence, time regions where violations occur. Put differently, it allows to identify, with a given confidence, intervals $I_{u,h} = [u-h, u+h]$ where m is not constant over time. It thus provides information on where the time trend is increasing/decreasing, which is important knowledge in many applications.

3.1 Construction of the test statistic

Throughout the section, we suppose that the trend m is continuously differentiable. The null hypothesis that m is constant can be formulated as $H_0 : m' = 0$, where m' denotes the first derivative of m. To construct a test statistic for the hypothesis H_0, we proceed analogously as in Section 2.1. To start with, we introduce the kernel averages

$$\widehat{\psi}'_T(u,h) = \sum_{t=1}^{T} w'_{t,T}(u,h) Y_t,$$

where the kernel weights $w'_{t,T}(u,h)$ are given by

$$w'_{t,T}(u,h) = \frac{\Lambda'_{t,T}(u,h)}{\{\sum_{t=1}^{T} \Lambda'_{t,T}(u,h)^2\}^{1/2}} \tag{3.1}$$

with

$$\Lambda'_{t,T}(u,h) = K\Big(\frac{\frac{t}{T}-u}{h}\Big)\Big[S_{T,0}(u,h)\Big(\frac{\frac{t}{T}-u}{h}\Big) - S_{T,1}(u,h)\Big].$$

Here, $S_{T,\ell}(u,h)$ is defined as in Section 2.1 and K is a kernel function which satisfies (C4). The kernel average $\widehat{\psi}'_T(u,h)$ is a rescaled version of the local linear estimator of the derivative $m'(u)$ with bandwidth h. Alternatively to the local linear weights defined in (3.1), we could employ the weights $w'_{t,T}(u,h) = K'(\frac{u-\frac{t}{T}}{h})/\{\sum_{t=1}^{T} K'(\frac{u-\frac{t}{T}}{h})^2\}^{1/2}$, where the kernel function K is assumed to be differentiable and K' is its derivative. To avoid boundary problems, we however work with the local linear weights from (3.1) throughout the paper. Our multiscale statistic is defined as

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

where $\lambda(h) = \sqrt{2\log\{1/(2h)\}}$ and the set $\mathcal{G}_T$ has been introduced in Section 2.1. As can be seen, the statistic $\widehat{\Psi}'_T$ is very similar to that from Section 2. Only the kernel

averages $\widehat{\psi}'_T(u,h)$ have a slightly different form.

3.2 The test procedure

As in Section 2.2, we define a Gaussian version Φ'_T of the test statistic $\widehat{\Psi}'_T$ under the null hypothesis H_0 by

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

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. Denoting the $(1-\alpha)$-quantile of Φ'_T by $q'_T(\alpha)$, our multiscale test of the hypothesis H_0: $m' = 0$ is 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

The theoretical analysis parallels that of Section 2.3. We first investigate the theoretical properties of the auxiliary statistic

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

where $\widehat{\phi}'_T(u,h) = \sum_{t=1}^T w'_{t,T}(u,h)\varepsilon_t$. The following result adapts Theorem 2.1 to our current test problem.

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

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

The proof of Theorem 3.1 is essentially the same as that of Theorem 2.1 and thus omitted. With the help of Theorem 3.1, we can derive the following theoretical properties of our multiscale test.

Proposition 3.2. *Let the conditions of Theorem 3.1 be satisfied.*

(a) Under the null hypothesis H_0, it holds that

$$\mathbb{P}\big(\widehat{\Psi}'_T \le q'_T(\alpha)\big) = (1-\alpha) + o(1).$$

(b) Consider any sequence of functions m_T with the following property: There exists $(u,h) \in \mathcal{G}_T$ with $[u-h, u+h] \subseteq [0,1]$ such that $m'_T(w) \ge c_T\sqrt{\log T/(Th^3)}$ for all

$w \in [u-h, u+h]$ *or* $-m'_T(w) \geq c_T\sqrt{\log T/(Th^3)}$ *for all* $w \in [u-h, u+h]$, *where* $\{c_T\}$ *is any sequence of positive numbers with* $c_T \to \infty$. *Then*

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

Part (a) of Proposition 3.2 is a simple consequence of Theorem 3.1. Part (b) can be proven by similar arguments as Proposition 2.3. The details are given in the Appendix. Taken together, the two parts of Proposition 3.2 show that our multiscale test has the correct (asymptotic) size and that it is able to detect certain local alternatives with probability tending to 1. We next consider the events

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

where the sets Π_T^+ and Π_T^- are given by

$$\Pi_T^+ = \big\{I_{u,h} = [u-h, u+h] : (u,h) \in \mathcal{A}_T^+\big\}$$
$$\Pi_T^- = \big\{I_{u,h} = [u-h, u+h] : (u,h) \in \mathcal{A}_T^-\big\}$$

with

$$\mathcal{A}_T^+ = \Big\{(u,h) \in \mathcal{G}_T : \frac{\widehat{\psi}'_T(u,h)}{\widehat{\sigma}} > q'_T(\alpha) + \lambda(h)\Big\}$$
$$\mathcal{A}_T^- = \Big\{(u,h) \in \mathcal{G}_T : -\frac{\widehat{\psi}'_T(u,h)}{\widehat{\sigma}} > q'_T(\alpha) + \lambda(h)\Big\}.$$

E_T^+ is the event that for each interval $I_{u,h} \in \Pi_T^+$, there is a subset $J_{u,h} \subseteq I_{u,h}$ with m being an increasing function on $J_{u,h}$. An analogous description applies to the event E_T^-. The following result shows that the events E_T^+ and E_T^- occur with asymptotic probability $\geq 1-\alpha$.

Proposition 3.3. *Under the conditions of Theorem 3.1, it holds that*

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

The proof of Proposition 3.3 parallels that of Proposition 2.4 and is thus omitted. As in Section 2.3, we can replace the sets Π_T^+ and Π_T^- in Proposition 3.3 by the corresponding sets of minimal intervals. The statement of Proposition 3.3 can be summarized as follows: With asymptotic probability $\geq 1-\alpha$, there is a subset $J_{u,h} \subseteq I_{u,h}$ for each interval $I_{u,h} \in \Pi_T^+$ such that m is an increasing function on $J_{u,h}$. Put differently, with

asymptotic probability $\geq 1-\alpha$, the trend m is increasing on some part of the interval $I_{u,h}$ for any $I_{u,h} \in \Pi_T^+$. An analogous statement holds for the intervals in the set Π_T^-. Our multiscale procedure thus allows us to identify, with a pre-specified confidence, time regions where there is an increase/decrease in the time trend m.

4 Testing for equality of time trends

In this section, we adapt the multiscale method developed in Section 2 to test the hypothesis that the trend functions in model (1.2) are all the same. More formally, we test the null hypothesis $H_0 : m_1 = m_2 = \ldots = m_n$ in model (1.2). As we will see, the proposed multiscale method does not only allow to test whether the null hypothesis is violated. It also provides information on where violations occur. More specifically, it allows to identify, with a pre-specified confidence, (i) trend functions which are different from each other and (ii) time intervals where these trend functions differ.

4.1 Construction of the test statistic

To start with, we introduce some notation. The i-th time series in model (1.2) satisfies the equation $Y_{it} = m_i(t/T) + \alpha_i + \varepsilon_{it}$, where ε_{it} are zero-mean error terms and α_i are (random or deterministic) intercepts. Defining $Y_{it}^\circ = Y_{it} - \alpha_i$, this equation can be rewritten as $Y_{it}^\circ = m_i(t/T) + \varepsilon_{it}$, which is a standard nonparametric regression equation. The variables Y_{it}° are not observed, but they can be easily approximated: As $\int_0^1 m_i(u)du = 0$ by normalization, the intercepts α_i can be estimated by $\widehat{\alpha}_i = T^{-1}\sum_{t=1}^T Y_{it}$. The variables $\widehat{Y}_{it} = Y_{it} - \widehat{\alpha}_i$ can thus be regarded as approximations of the unknown quantities Y_{it}°. We further let $\widehat{\sigma}_i^2$ be an estimator of the long-run error variance $\sigma_i^2 = \sum_{\ell=-\infty}^{\infty}\gamma_i(\ell)$ with $\gamma_i(\ell) = \mathrm{Cov}(\varepsilon_{i0}, \varepsilon_{i\ell})$ and assume that $\widehat{\sigma}_i^2 = \sigma_i^2 + o_p(\rho_T)$ with $\rho_T = T^{1/q}/\sqrt{Th_{\min}\log T}$. Details on how to construct estimators of σ_i^2 are deferred to Section 5. To keep the exposition simple, we assume that $\sigma_i^2 = \sigma^2$ for all i and set $\widehat{\sigma}^2 = n^{-1}\sum_{i=1}^n \widehat{\sigma}_i^2$ in what follows. It is not difficult to adapt our methods and theory to the case where the variances σ_i^2 differ across i.

We are now ready to introduce the multiscale statistic for testing the hypothesis $H_0 : m_1 = m_2 = \ldots = m_n$. For any pair of time series i and j, we define the kernel averages

$$\widehat{\psi}_{ij,T}(u,h) = \sum_{t=1}^T w_{t,T}(u,h)(\widehat{Y}_{it} - \widehat{Y}_{jt}),$$

where the kernel weights are defined as in (2.1). The kernel average $\widehat{\psi}_{ij,T}(u,h)$ can be regarded as measuring the distance between the two trend curves m_i and m_j on the interval $[u-h, u+h]$. Similar as in Section 2.1, we aggregate the kernel averages

$\widehat{\psi}_{ij,T}(u,h)$ for all $(u,h) \in \mathcal{G}_T$ by the multiscale statistic

$$\widehat{\Psi}_{ij,T} = \max_{(u,h)\in\mathcal{G}_T} \Big\{ \Big| \frac{\widehat{\psi}_{ij,T}(u,h)}{\sqrt{2}\widehat{\sigma}} \Big| - \lambda(h) \Big\},$$

where $\lambda(h) = \sqrt{2\log\{1/(2h)\}}$ and the set $\mathcal{G}_T$ has been introduced in Section 2.1. The statistic $\widehat{\Psi}_{ij,T}$ can be interpreted as some sort of distance measure between the two curves m_i and m_j. We finally define the multiscale statistic for testing the null hypothesis $H_0 : m_1 = m_2 = \ldots = m_n$ as

$$\widehat{\Psi}_{n,T} = \max_{1\le i<j\le n} \widehat{\Psi}_{ij,T},$$

that is, we define it as the maximal distance $\widehat{\Psi}_{ij,T}$ between any pair of curves m_i and m_j with $i \neq j$.

4.2 The test procedure

Let Z_{it} for $1 \le t \le T$ and $1 \le i \le n$ be independent standard normal random variables. For each i and j, define the Gaussian statistic

$$\Phi_{ij,T} = \max_{(u,h)\in\mathcal{G}_T} \Big\{ \Big| \frac{\phi_{ij,T}(u,h)}{\sqrt{2}\sigma} \Big| - \lambda(h) \Big\},$$

where $\phi_{ij,T}(u,h) = \sum_{t=1}^{T} w_{t,T}(u,h)\,\sigma(Z_{it} - Z_{jt})$. Moreover, define the statistic

$$\Phi_{n,T} = \max_{1\le i<j\le n} \Phi_{ij,T}$$

and denote its $(1-\alpha)$-quantile by $q_{n,T}(\alpha)$. Our multiscale test of the hypothesis $H_0 : m_1 = m_2 = \ldots = m_n$ is defined as follows: For a given significance level $\alpha \in (0,1)$, we reject H_0 if $\widehat{\Psi}_{n,T} > q_{n,T}(\alpha)$.

4.3 Theoretical properties of the test

Similar as in the previous sections, we introduce the auxiliary statistic

$$\widehat{\Phi}_{n,T} = \max_{1\le i<j\le n} \widehat{\Phi}_{ij,T},$$

where

$$\widehat{\Phi}_{ij,T} = \max_{(u,h)\in\mathcal{G}_T} \Big\{ \Big| \frac{\widehat{\phi}_{ij,T}(u,h)}{\sqrt{2}\widehat{\sigma}^{\circ}} \Big| - \lambda(h) \Big\}$$

and $\widehat{\phi}_{ij,T}(u,h) = \sum_{t=1}^{T} w_{t,T}(u,h)(\varepsilon_{it} - \varepsilon_{jt})$. Here, $\widehat{\sigma}^{\circ}$ is the same estimator as $\widehat{\sigma}$ with $\widehat{Y}_{it}$ replaced by Y_{it}° for $1 \le t \le T$. Since $\widehat{\sigma} = \sigma + o_p(\rho_T)$ with $\rho_T = T^{1/q}/\sqrt{Th_{\min}\log T}$,

we can expect that $\widehat{\sigma}^\circ = \sigma + o_p(\rho_T)$ as well. This is indeed true for most estimators of σ. When using the difference-based estimation methods from Section 5, we even have that $\widehat{\sigma}^\circ = \widehat{\sigma}$. In the sequel, we make the general assumption that $\widehat{\sigma}$ and $\widehat{\sigma}^\circ$ are any estimators with the property that $\widehat{\sigma} = \sigma + o_p(\rho_T)$ and $\widehat{\sigma}^\circ = \sigma + o_p(\rho_T)$. Our first theoretical result characterizes the asymptotic behaviour of the statistic $\widehat{\Phi}_{n,T}$ and parallels Theorem 2.1 from Section 2.

Theorem 4.1. *Suppose that the error processes $\mathcal{E}_i = \{\varepsilon_{it} : 1 \leq t \leq T\}$ are independent across i and satisfy (C1)–(C3) for each i. Moreover, let (C4)–(C6) be fulfilled. Then*

$$\mathbb{P}\big(\widehat{\Phi}_{n,T} \leq q_{n,T}(\alpha)\big) = (1-\alpha) + o(1).$$

Theorem 4.1 can be proven by slightly modifying the arguments for Theorem 2.1. The details are provided in the Appendix. With the help of Theorem 4.1, we can derive the following theoretical properties of our multiscale test.

Proposition 4.2. *Let the conditions of Theorem 4.1 be satisfied.*

(a) Under the null hypothesis $H_0 : m_1 = m_2 = \ldots = m_n$, it holds that

$$\mathbb{P}\big(\widehat{\Psi}_{n,T} \leq q_{n,T}(\alpha)\big) = (1-\alpha) + o(1).$$

(b) Assume that for some indices i and j, the functions $m_{i,T}$ and $m_{j,T}$ have the following property: There exists $(u,h) \in \mathcal{G}_T$ with $[u-h, u+h] \subseteq [0,1]$ such that $m_{i,T}(w) - m_{j,T}(w) \geq c_T\sqrt{\log T/(Th)}$ for all $w \in [u-h, u+h]$ or $m_{j,T}(w) - m_{i,T}(w) \geq c_T\sqrt{\log T/(Th)}$ for all $w \in [u-h, u+h]$, where $\{c_T\}$ is any sequence of positive numbers with $c_T \to \infty$. Then

$$\mathbb{P}\big(\widehat{\Psi}_{n,T} \leq q_{n,T}(\alpha)\big) = o(1).$$

Part (a) of Proposition 4.2 is a straightforward consequence of Theorem 4.1. The proof of part (b) is very similar to that of Proposition 2.3 and thus omitted.

4.4 Clustering of time trends

Consider a situation in which the null hypothesis $H_0 : m_1 = m_2 = \ldots = m_n$ is violated. Even though some of the trend functions are different in this case, part of them may still be the same. Put differently, there may be groups of time series which have the same time trend. Formally speaking, we define a group structure as follows: There exist sets or groups of time series $G_1, \ldots, G_N$ with $N \leq n$ and $\{1, \ldots, n\} = \dot{\bigcup}_{\ell=1}^{N} G_\ell$ such that for each $1 \leq \ell \leq N$,

$$m_i = g_\ell \quad \text{for all } i \in G_\ell,$$

where g_ℓ are group-specific trend functions. Hence, the time series which belong to the group G_ℓ all have the same time trend g_ℓ. The group-specific trend functions g_ℓ are of course supposed to be different across groups G_ℓ. More specifically, for any $\ell \neq \ell'$, the trends $g_\ell = g_{\ell,T}$ and $g_{\ell'} = g_{\ell',T}$ are supposed to differ in the following sense: There exists $(u,h) \in \mathcal{G}_T$ with $[u-h, u+h] \subseteq [0,1]$ such that $g_{\ell,T}(w) - g_{\ell',T}(w) \geq c_T\sqrt{\log T/(Th)}$ for all $w \in [u-h, u+h]$ or $g_{\ell',T}(w) - g_{\ell,T}(w) \geq c_T\sqrt{\log T/(Th)}$ for all $w \in [u-h, u+h]$, where $0 < c_T \to \infty$.

In many applications, it is natural to suppose that there is a group structure in the data as defined above. In this case, a particular interest lies in estimating the unknown groups from the data sample at hand. In what follows, we combine our multiscale methods with a clustering algorithm to achieve this. More specifically, we use the multiscale statistics $\widehat{\Psi}_{ij,T}$ as distance measures which are fed into a hierarchical clustering algorithm. To describe the algorithm, we first need to introduce the notion of a dissimilarity measure: Let $S \subseteq \{1,\dots,n\}$ and $S' \subseteq \{1,\dots,n\}$ be two sets of time series from our sample. We define a dissimilarity measure between S and S' by setting

$$\widehat{\Delta}(S,S') = \max_{\substack{i\in S,\\ j\in S'}} \widehat{\Psi}_{ij,T}. \tag{4.1}$$

This is commonly called a complete linkage measure of dissimilarity. Alternatively, we may work with an average or a single linkage measure. We now combine the dissimilarity measure $\widehat{\Delta}$ with a hierarchical agglomerative clustering (HAC) algorithm which proceeds as follows:

Step 0 *(Initialization):* Let $\widehat{G}_i^{[0]} = \{i\}$ denote the i-th singleton cluster for $1 \leq i \leq n$ and define $\{\widehat{G}_1^{[0]},\dots,\widehat{G}_n^{[0]}\}$ to be the initial partition of subjects into clusters.

Step r (Iteration): Let $\widehat{G}_1^{[r-1]},\dots,\widehat{G}_{n-(r-1)}^{[r-1]}$ be the $n-(r-1)$ clusters from the previous step. Determine the pair of clusters $\widehat{G}_\ell^{[r-1]}$ and $\widehat{G}_{\ell'}^{[r-1]}$ for which

$$\widehat{\Delta}(\widehat{G}_\ell^{[r-1]}, \widehat{G}_{\ell'}^{[r-1]}) = \min_{1\leq k<k'\leq n-(r-1)} \widehat{\Delta}(\widehat{G}_k^{[r-1]}, \widehat{G}_{k'}^{[r-1]})$$

and merge them into a new cluster.

Iterating this procedure for $r = 1,\dots,n-1$ yields a tree of nested partitions $\{\widehat{G}_1^{[r]},\dots$ $\dots,\widehat{G}_{n-r}^{[r]}\}$, which can be graphically represented by a dendrogram. Roughly speaking, the HAC algorithm merges the n singleton clusters $\widehat{G}_i^{[0]} = \{i\}$ step by step until we end up with the cluster $\{1,\dots,n\}$. In each step of the algorithm, the closest two clusters are merged, where the distance between clusters is measured in terms of the dissimilarity $\widehat{\Delta}$. We refer the reader to Section 14.3.12 in Hastie et al. (2009) for an overview of hierarchical clustering methods.

When the number of groups N is known, we estimate the group structure $\{G_1,\dots,G_N\}$

by the N-partition $\{\widehat{G}_1^{[n-N]}, \ldots, \widehat{G}_N^{[n-N]}\}$ produced by the HAC algorithm. When N is unknown, we estimate it by the $\widehat{N}$-partition $\{\widehat{G}_1^{[n-\widehat{N}]}, \ldots, \widehat{G}_{\widehat{N}}^{[n-\widehat{N}]}\}$, where $\widehat{N}$ is an estimator of N. The latter is defined as

$$\widehat{N} = \min\Big\{r = 1, 2, \ldots \,\Big|\, \max_{1\leq \ell \leq r} \widehat{\Delta}\big(\widehat{G}_\ell^{[n-r]}\big) \leq q_{n,T}(\alpha)\Big\},$$

where we write $\widehat{\Delta}(S) = \widehat{\Delta}(S, S)$ for short and $q_{n,T}(\alpha)$ is the $(1-\alpha)$-quantile of $\Phi_{n,T}$ defined in Section 4.2.

The following proposition summarizes the theoretical properties of the estimators $\widehat{N}$ and $\{\widehat{G}_1, \ldots, \widehat{G}_{\widehat{N}}\}$, where we use the shorthand $\widehat{G}_\ell = \widehat{G}_\ell^{[n-\widehat{N}]}$ for $1 \leq \ell \leq \widehat{N}$.

Proposition 4.3. *Let the conditions of Theorem 4.1 be satisfied. Then*

$$\mathbb{P}\Big(\big\{\widehat{G}_1, \ldots, \widehat{G}_{\widehat{N}}\big\} = \{G_1, \ldots, G_N\}\Big) \geq (1-\alpha) + o(1)$$

and

$$\mathbb{P}\big(\widehat{N} = N\big) \geq (1-\alpha) + o(1).$$

This result allows us to make statistical confidence statements about the estimated clusters $\{\widehat{G}_1, \ldots, \widehat{G}_{\widehat{N}}\}$ and their number $\widehat{N}$. In particular, we can claim with asymptotic confidence $\geq 1-\alpha$ that the estimated group structure is identical to the true group structure. Note that it is possible to let the significance level α depend on the sample size T in Proposition 4.3. In particular, we can allow $\alpha = \alpha_T$ to converge slowly to zero as $T \to \infty$, in which case we obtain that $\mathbb{P}(\{\widehat{G}_1, \ldots, \widehat{G}_{\widehat{N}}\} = \{G_1, \ldots, G_N\}) \to 1$ and $\mathbb{P}(\widehat{N} = N) \to 1$. The proof of Proposition 4.3 can be found in the Appendix.

Our multiscale methods do not only allow to compute estimators of the unknown groups $G_1, \ldots, G_N$. They also provide information on the locations where two group-specific trend functions g_ℓ and $g_{\ell'}$ differ from each other. To turn this claim into a mathematically precise statement, we need to introduce some notation. First of all, note that the indexing of the estimators $\widehat{G}_1, \ldots, \widehat{G}_{\widehat{N}}$ is completely arbitrary. We could, for example, change the indexing according to the rule $\ell \mapsto \widehat{N} - \ell + 1$. In what follows, we suppose that the estimated groups are indexed such that $P(\widehat{G}_\ell = G_\ell) \geq (1-\alpha) + o(1)$ for all ℓ. Theorem 4.3 implies that this is possible without loss of generality. Keeping this convention in mind, we define the sets

$$\mathcal{A}_{n,T}(\ell, \ell') = \Big\{(u,h) \in \mathcal{G}_T : \Big|\frac{\widehat{\psi}_{ij,T}(u,h)}{\widehat{\sigma}}\Big| > q_{n,T}(\alpha) + \lambda(h) \text{ for some } i \in \widehat{G}_\ell, j \in \widehat{G}_{\ell'}\Big\}$$

and

$$\Pi_{n,T}(\ell, \ell') = \big\{I_{u,h} = [u-h, u+h] : (u,h) \in \mathcal{A}_{n,T}(\ell, \ell')\big\}$$

for $1 \leq \ell < \ell' \leq \widehat{N}$. An interval $I_{u,h}$ is contained in $\Pi_{n,T}(\ell, \ell')$ if our multiscale test indicates a significant difference between the trends m_i and m_j on the interval $I_{u,h}$

for some $i \in \widehat{G}_\ell$ and $j \in \widehat{G}_{\ell'}$. Put differently, $I_{u,h} \in \Pi_{n,T}(\ell, \ell')$ if the test suggests a significant difference between the trends of the ℓ-th and the ℓ'-th group on the interval $I_{u,h}$. We further let

$$E_{n,T}(\ell, \ell') = \Big\{ \forall I_{u,h} \in \Pi_{n,T}(\ell, \ell') : g_\ell(v) \neq g_{\ell'}(v) \text{ for some } v \in I_{u,h} = [u-h, u+h] \Big\}$$

be the event that the group-specific time trends g_ℓ and $g_{\ell'}$ differ on all intervals $I_{u,h} \in \Pi_{n,T}(\ell, \ell')$. With this notation at hand, we can make the following formal statement whose proof is given in the Appendix.

Proposition 4.4. *Under the conditions of Proposition 4.3, the event*

$$E_{n,T} = \Big\{ \bigcap_{1 \leq \ell < \ell' \leq \widehat{N}} E_{n,T}(\ell, \ell') \Big\} \cap \Big\{ \widehat{N} = N \text{ and } \widehat{G}_\ell = G_\ell \text{ for all } \ell \Big\}$$

asymptotically occurs with probability $\geq 1 - \alpha$*, that is,*

$$\mathbb{P}\big(E_{n,T}\big) \geq (1 - \alpha) + o(1).$$

The statement of Proposition 4.4 remains to hold true when the sets of intervals $\Pi_{n,T}(\ell, \ell')$ are replaced by the corresponding sets of minimal intervals. According to Proposition 4.4, the sets $\Pi_{n,T}(\ell, \ell')$ allow us to locate, with a pre-specified confidence, time regions where the group-specific trend functions g_ℓ and $g_{\ell'}$ differ from each other. In particular, we can claim with asymptotic confidence $\geq 1-\alpha$ that the trend functions g_ℓ and g'_ℓ differ on all intervals $I_{u,h} \in \Pi_{n,T}(\ell, \ell')$.

5 Estimation of the long-run error variance

We now discuss how to estimate the long-run error variance $\sigma^2 = \sum_{\ell=-\infty}^{\infty} \gamma(\ell)$ with $\gamma(\ell) = \mathrm{Cov}(\varepsilon_0, \varepsilon_\ell)$ in model (1.1). The same methods can be applied in the context of model (1.2). A number of different methods have been established in the literature to estimate the long-run error variance σ^2 in the trend model (1.1) under various assumptions on the error terms. In what follows, we give a brief overview of estimation methods which are suitable for our purposes. We in particular focus attention on difference-based methods as these have the following advantage: They do not involve a nonparametric estimator of the function m and thus do not require to specify a smoothing parameter for the estimation of m.

In principle, it is possible to construct an estimator of σ^2 under the general conditions on the error process laid out in Section 1 (or at least under somewhat stronger versions of these conditions). However, as is well-known, it is quite involved to estimate the long-run variance of a time series process under general conditions, the resulting

estimators often tending to be quite imprecise. From a practical point of view, one might thus prefer to impose some time series model on the error terms and to estimate σ^2 under the restrictions of this model. Of course, this will create some bias due to misspecification. However, as long as the model gives a reasonable approximation to the true error process, this bias may very well be less severe than the error stemming from the instable behaviour of a general estimator of σ^2. In what follows, we consider an autoregressive (AR) model for the error terms since this error model is widely used in practice and is in particular appropriate for our applications in Section ??.

5.1 Independent error terms

Before we discuss the case of autoregressive error terms, we introduce the idea of difference-based methods for estimating σ^2 in the simple case of i.i.d. errors ε_t. In this case, σ^2 is identical to the variance of the random variables ε_t, that is, $\sigma^2 = \mathrm{Var}(\varepsilon_t)$. Let $D_\ell Y_t = Y_t - Y_{t-\ell}$ denote the difference between Y_t and $Y_{t-\ell}$ and suppose that m is sufficiently smooth. In particular, assume that m is Lipschitz continuous on $[0,1]$, that is, $|m(u) - m(v)| \leq C|u-v|$ for all $u, v \in [0,1]$ and some constant $C < \infty$. Under these conditions, it holds that $|m(t/T) - m(\{t-\ell\}/T)| \leq C\ell/T$, which implies that $D_\ell Y_t = D_\ell \varepsilon_t + O(\ell/T)$ uniformly over t. Hence, the observed differences $D_\ell Y_t$ approximate the unobserved differences of the error terms $D_\ell \varepsilon_t$. This together with the fact that $\mathbb{E}[\{D_\ell \varepsilon_t\}^2]/2 = \sigma^2$ suggests to estimate σ^2 by

$$\widehat{\sigma}^2 = \frac{1}{T-\ell} \sum_{t=\ell+1}^{T} \{D_\ell Y_t\}^2/2,$$

where most commonly $\ell = 1$. As can be easily verified, the estimator $\widehat{\sigma}^2$ has the property that $\widehat{\sigma}^2 = \sigma^2 + O_p(T^{-1/2})$.

5.2 Autoregressive error terms

The differencing approach presented above can be extended to more complicated error structures. For the case of k-dependent error terms, estimators have been proposed by Müller and Stadtmüller (1988), Herrmann et al. (1992) and Tecuapetla-Gómez and Munk (2017) among others. We here focus attention to the case of autoregressive error terms. Specifically, we suppose that $\{\varepsilon_t\}$ is an AR(p) process of the form

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

where $a_1, \ldots, a_p$ are unknown parameters and η_t are i.i.d. innovations with $\mathbb{E}[\eta_t] = 0$ and $\mathbb{E}[\eta_t^2] = \sigma_\eta^2$. Throughout the discussion, we assume that $\{\varepsilon_t\}$ is a stationary

and causal AR(p) process of known order p with finite fourth moment $\mathbb{E}[\varepsilon_t^4] < \infty$. A difference-based method to estimate the long-run variance σ^2 of the AR(p) error process $\{\varepsilon_t\}$ in model (1.1) has been developed in Hall and Van Keilegom (2003). Their estimator $\widehat{\sigma}^2$ is constructed in the following three steps:

Step 1. We first set up an estimator of the autocovariance $\gamma(\ell) = \mathrm{Cov}(\varepsilon_t, \varepsilon_{t+\ell})$ for a given lag ℓ. As in the case of independent errors, it holds that $D_\ell Y_t = D_\ell \varepsilon_t + O(\ell/T)$ uniformly over t provided that m is Lipschitz. This together with the fact that $\mathbb{E}[\{D_\ell \varepsilon_t\}^2]/2 = \gamma(0) - \gamma(\ell)$ motivates to estimate $\gamma(0)$ by

$$\widehat{\gamma}(0) = \frac{1}{L_2 - L_1 + 1} \sum_{r=L_1}^{L_2} \frac{1}{2(T-r)} \sum_{t=r+1}^{T} \{D_r Y_t\}^2,$$

where $L_1 \leq L_2$ are tuning parameters which are discussed in more detail below. Moreover, an estimator of $\gamma(\ell)$ for $1 \leq \ell \leq p$ is given by

$$\widehat{\gamma}(\ell) = \widehat{\gamma}(0) - \frac{1}{2(T-\ell)} \sum_{t=\ell+1}^{T} \{D_\ell Y_t\}^2.$$

As $\gamma(\ell) = \gamma(-\ell)$, we finally set $\widehat{\gamma}(-\ell) = \widehat{\gamma}(\ell)$ for $1 \leq \ell \leq p$.

Step 2. We next estimate the AR coefficients $(a_1, \ldots, a_p)^\top$ by the Yule-Walker estimators $(\widehat{a}_1, \ldots, \widehat{a}_p)^\top = \widehat{\Gamma}^{-1}(\widehat{\gamma}(1), \ldots, \widehat{\gamma}(p))^\top$, where the matrix $\widehat{\Gamma}$ is given by $\widehat{\Gamma} = \{\widehat{\gamma}(|k-\ell|)\}_{1 \leq k, \ell \leq p}$.

Step 3. Let $\widehat{d}_0 = 1$ and define the parameters $\widehat{d}_1, \widehat{d}_2, \ldots$ by the equation

$$1 + \sum_{\ell=1}^{\infty} \widehat{d}_\ell z^\ell = \Big(1 - \sum_{j=1}^{p} \widehat{a}_j z^j\Big)^{-1}.$$

In the AR(1) case $\varepsilon_t = a\varepsilon_{t-1} + \eta_t$, for instance, it holds that $\sum_{\ell=0}^{\infty} \widehat{a}^\ell z^\ell = (1 - \widehat{a}z)^{-1}$ and thus $\widehat{d}_\ell = \widehat{a}^\ell$ for $\ell \geq 1$. The variance $\sigma_\eta^2 = \mathbb{E}[\eta_t^2]$ of the innovations can be estimated by $\widehat{\sigma}_\eta^2 = \widehat{\gamma}(0)/(\sum_{\ell=0}^{\infty} \widehat{d}_\ell^2)$. With this notation at hand, the estimator $\widehat{\sigma}^2$ of the long-run variance σ^2 is defined as

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

The estimator $\widehat{\sigma}^2$ depends on the two tuning parameters L_1 and L_2 which are required to compute $\widehat{\gamma}(0)$. To better understand the role of these tuning parameters, let us have a closer look at the estimator $\widehat{\gamma}(0)$. As $\mathbb{E}[\{D_\ell Y_t\}^2]/2 = \mathbb{E}[\{D_\ell \varepsilon_t\}^2]/2 + O(\{\ell/T\}^2) =$

$\gamma(0) - \gamma(\ell) + O(\{\ell/T\}^2)$, it can be easily shown that

$$\mathbb{E}\big[\widehat{\gamma}(0)\big] = \gamma(0) + \frac{1}{L_2 - L_1 + 1} \sum_{r=L_1}^{L_2} \gamma(r) + O\Big(\Big\{\frac{L_2}{T}\Big\}^2\Big).$$

Since $\{\varepsilon_t\}$ is an AR(p) process, the autocovariances $\gamma(r)$ decay exponentially fast to zero as $r \to \infty$. Hence, the bias term $\sum_{r=L_1}^{L_2} \gamma(r)/(L_2 - L_1 + 1)$ is asymptotically negligible if L_1 grows sufficiently fast with the sample size T. Due to the exponential decay of the autocovariances, it in particular suffices to assume that $L_1/\log T \to \infty$. For the second bias term $O(\{L_2/T\}^2)$ to be asymptotically negligible, we need to assume that L_2 grows more slowly than the sample size T. In practice, L_1 should be chosen so large that the autocovariances $\gamma(\ell)$ with $\ell \geq L_1$ can be expected to be close to zero, ensuring that the bias term $\sum_{r=L_1}^{L_2} \gamma(r)/(L_2 - L_1 + 1)$ is sufficiently small. The choice of L_2 can be expected to be less important in practice than that of L_1 as long as we do not pick L_2 too close to the sample size T. As pointed out in Hall and Van Keilegom (2003), it can be shown that $\widehat{\sigma}^2 = \sigma^2 + O_p(T^{-1/2})$ provided that $L_1/\log T \to \infty$ and $L_2 = O(T^{1/2})$.

6 Simulations

In this section we present a simulation study to assess the performance of our baseline test procedure from Section 1.

Our statistic $\widehat{\Psi}_T$ depends not only on the sample size T, but also on the choice of the kernel function $K(x)$ and the set $\mathcal{G}_T$. Throughout our numerical experiments, the Epanechnikov kernel $K(x) = \mathbb{1}_{|x|\leq 1}\frac{3}{4}(1 - x^2)$ and the set $\mathcal{G}_T = \{(u, h) : u = (5k)/T$ and $h = (3 + 5l)/T$ for some $k, l \in \mathbb{N}, k \leq T/5, l \leq T/20\}$ are used. The Epanechnikov kernel $K(x)$ satisfies the condition (C4), whereas the set $\mathcal{G}_T$ with $|\mathcal{G}_T| = O(T^2)$ satisfies the conditions (C5)-(C6).

6.1 Size of the test

We first provide some evidence on the size of the test. Data was generated by the model 1.1 under the null hypotheis $H_0 : m = 0$ with AR(1) errors

$$\varepsilon_t = a\varepsilon_{t-1} + \eta_t.$$

Here η_t are i.i.d. Gaussian inovations with $\mathbb{E}[\eta_t] = 0$ and $\mathbb{E}[\eta_t^2] = \sigma_\eta^2$. As was already stated before, such process satisfies the conditions (C1) - (C3), therefore, 2.1 holds true.

We choose parameters a and σ_η^2 such that the generated data resembles the real-life

temperature dataset used for illustrating our method in Section 7. In order to do that, we assume that the error structure in the real-life data also follows an AR(1) process and we estimate AR coefficient a_1 and the variance of the innovation σ_η^2 using the method described in Section 5.2. We get the results $\hat{a}_1 = 0.5$ and $\hat{\sigma}_\eta^2 = 0.6$ which we use for generating errors $\{\varepsilon_t\}$.

Table 1 shows size computations for various sample size ($T = 250, 350, 500, 1000$) and confidence levels ($\alpha = 0.01, 0.05, 0.10$). The simulations are based on 1000 replications and give results for one-sided tests under the null hypothesis $H_0 : m = 0$. As can be seen from the table 1, our test has approximately correct size even for small values of T and the accuracy does not decrease with the sample size.

Table 1: Size of the test

	0.01	0.05	0.1
250	0.007	0.033	0.072
350	0.005	0.041	0.087
500	0.015	0.054	0.078

6.2 Power of the test

Tables 2 - 5 show power computations for various sample size ($T = 250, 350, 500, 1000$), confidence levels ($\alpha = 0.01, 0.05, 0.10$) and under different alternatives H_1. The simulations are based on 1000 replications and give results for one-sided tests under the alternative hypothesis $H_1 : m \neq 0$. As can be seen from the tables 2 - 5, our test has approximately correct size even for small values of T and the accuracy does not decrease with the sample size.

Power of the test. Each table corresponds to different value of a, the endpoint of the trend function $m(\cdot)$.

Table 2: $a = 0.25$

	0.01	0.05	0.1
250	0.032	0.092	0.155
350	0.018	0.090	0.199
500	0.069	0.169	0.212

Table 3: $a = 0.50$

	0.01	0.05	0.1
250	0.153	0.265	0.399
350	0.192	0.333	0.517
500	0.355	0.571	0.694

Table 4: $a = 0.65$

	0.01	0.05	0.1
250	0.321	0.471	0.590
350	0.416	0.620	0.744
500	0.686	0.846	0.900

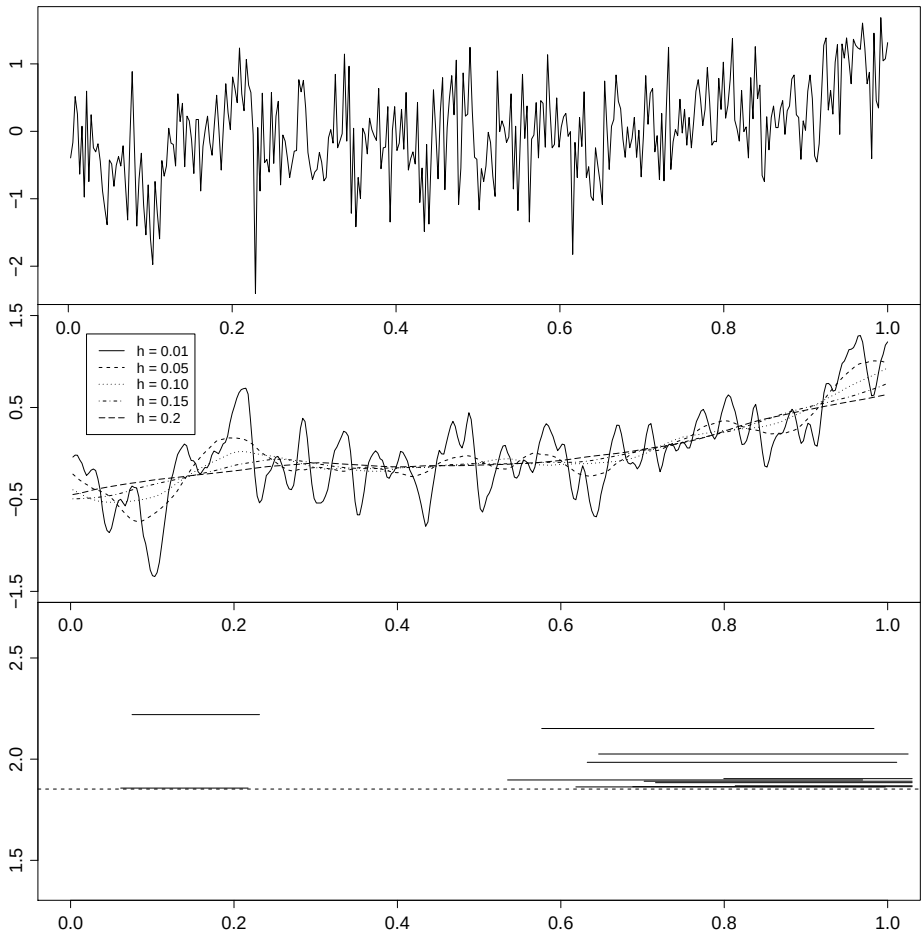
Table 5: $a = 0.75$

	0.01	0.05	0.1
250	0.461	0.625	0.754
350	0.571	0.793	0.888
500	0.862	0.936	0.968

7 Data analysis

In this section, our methodology is applied to the problem of analyzing the temperature trends. For our analysis, we use climate data collected by The Met Office, a United Kingdom national weather service. The dataset of mean yearly temperature used to illustrate methodology from Section 1 consists of 359 observation of mean yearly temperature in England that covers years 1768 - 2017. The dataset of mean monthly temperature for various stations in the United Kingdom consists of 303 observations.

Figure 1: Yearly temperature data for England



Appendix

In what follows, we prove the main theoretical results of the paper. Throughout the Appendix, the symbol C denotes a universal real constant which may take a different value on each occurrence. We use the following notation: 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 (2.4) 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 2.1, 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)*

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

where Φ_T is a Gaussian statistic as defined in (2.3).

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 let

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

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 define

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

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

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

$$\big|\widetilde{\Phi}_T - \Phi_T\big| \leq \big|\widetilde{\Phi}_T - \Phi_T^*\big| + \big|\Phi_T^* - \Phi_T\big| = \big|\widetilde{\Phi}_T - \Phi_T^*\big| + o_p\Big(\frac{T^{1/q}}{\sqrt{Th_{\min}}}\Big), \tag{A.2}$$

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 $\widetilde{\sigma}^2 = \sigma^2 + o_p(T^{1/q}/\sqrt{Th_{\min}\log T})$. Straightforward calculations yield that

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

Using summation by parts, we further obtain that

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

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

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

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 (2.3). 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. *Set $\delta_T = T^{1/q}/\sqrt{Th_{\min}}$. Under the conditions of Theorem 2.1, it holds that*

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

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 $\overline{\sigma}$.

For the sake of completeness, the proof of Proposition A.3 is provided at the end of the Appendix. 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 $\overline{\mu} = \max_{1\leq j\leq p} |\mu_j| \leq C\sqrt{\log T}$, and (ii) $\sigma_j^2 := \mathrm{Var}(X_j) = 1$ for all j. Since $\sigma_j = 1$ for all j, it holds that $a_p = b_p$. Moreover, as the variables $(X_j - \mu_j)/\sigma_j$ are standard normal, we have that $a_p = b_p \leq \sqrt{2\log(2p)} \leq C\sqrt{\log T}$. With this notation at hand, we can apply Proposition A.3 to obtain that

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

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

Proof of Theorem 2.1

To prove Theorem 2.1, 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

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

where Φ_T is a Gaussian statistic as defined in (2.3). 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 2.1. 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 $\delta_T = o(1)$. 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

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

□

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

that

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

Note that with the help of Theorem 2.1 in Dümbgen and Spokoiny (2001), we can further show that $\Phi_T = O_p(1)$. Together with (A.8), this says that the Gaussian multiscale statistic Φ_T is asymptotically tight and does not concentrate too strongly in small regions of the form $[x - \delta_T, x + \delta_T]$. Putting everything together, we are now in a position to apply Lemma A.4, which in turn yields (A.5). This completes the proof of Theorem 2.1.

Proof of Proposition 2.3

Define $\widehat{\psi}_T^A(u,h) = \sum_{t=1}^T w_{t,T}(u,h)\varepsilon_t$, $\widehat{\psi}_T^B(u,h) = \sum_{t=1}^T w_{t,T}(u,h) m_T(t/T)$ and

$$\widehat{\psi}_T^*(u,h) = \frac{\sqrt{Th}\int_0^1 h^{-1}K(\frac{w-u}{h})[S_2(u,h) - S_1(u,h)(\frac{w-u}{h})]m_T(w)dw}{\{\int_0^1 h^{-1}K^2(\frac{w-u}{h})[S_2(u,h) - S_1(u,h)(\frac{w-u}{h})]^2 dw\}^{1/2}},$$

where $S_\ell(u,h) = \int_0^1 h^{-1}K(\frac{w-u}{h})(\frac{w-u}{h})^\ell dw$. With the help of Proposition A.1, we can show that

$$\max_{(u,h)\in\mathcal{G}_T} |\widehat{\psi}_T^A(u,h)| = O_p(\sqrt{\log T}). \tag{A.9}$$

Moreover, standard calculations yield that

$$\big|\widehat{\psi}_T^B(u,h) - \widehat{\psi}_T^*(u,h)\big| \le \frac{C}{Th}, \tag{A.10}$$

where the constant C is independent of u, h and T. By assumption, there exists $(u,h) \in \mathcal{G}_T$ with $[u-h, u+h] \subseteq [0,1]$ such that $m_T(w) \ge c_T\sqrt{\log T/(Th)}$ for all $w \in [u-h, u+h]$. For this (u,h), it holds that

$$\widehat{\psi}_T^*(u,h) = \kappa^{-1}\sqrt{Th}\int_0^1 h^{-1}K\Big(\frac{w-u}{h}\Big)m_T(w)dw \ge \kappa^{-1}c_T\sqrt{\log T}, \tag{A.11}$$

where $\kappa = \int K^2(\varphi)d\varphi$. Using (A.9)–(A.11) and noticing that $\lambda(h) \le \lambda(h_{\min}) \le C\sqrt{\log T}$, we obtain that

$$\begin{aligned}\widehat{\Psi}_T &\ge \max_{(u,h)\in\mathcal{G}_T} \frac{|\widehat{\psi}_T^B(u,h)|}{\widehat{\sigma}} - \max_{(u,h)\in\mathcal{G}_T}\Big\{\frac{|\widehat{\psi}_T^A(u,h)|}{\widehat{\sigma}} + \lambda(h)\Big\} \\ &= \max_{(u,h)\in\mathcal{G}_T} \frac{|\widehat{\psi}_T^B(u,h)|}{\widehat{\sigma}} + O_p(\sqrt{\log T}) \ge \frac{c_T\sqrt{\log T}}{\kappa\widehat{\sigma}} + O_p(\sqrt{\log T}).\end{aligned} \tag{A.12}$$

Since $q_T(\alpha) = O(\sqrt{\log T})$ for any fixed $\alpha \in (0,1)$, (A.12) immediately implies that $\mathbb{P}(\widehat{\Psi}_T \le q_T(\alpha)) = o(1)$.

Proof of Proposition 2.4

The statement of Proposition 2.4 is a consequence of the following observation: For all $(u,h) \in \mathcal{G}_T$ with

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

it holds that $\mathbb{E}[\widehat{\psi}_T(u,h)] \neq 0$, which in turn implies that $m(v) \neq 0$ for some $v \in I_{u,h}$. From this observation, we can infer the following: 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 $(u,h) \in \mathcal{A}_T$, $m(v) \neq 0$ for some $v \in I_{u,h}$. Hence, we obtain that

$$\big\{\widehat{\Phi}_T \leq q_T(\alpha)\big\} \subseteq E_T.$$

As a result, we arrive at

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

where the last equality holds by Theorem 2.1.

Proof of Proposition 3.2

We only need to prove part (b). By assumption, there exists $(u,h) \in \mathcal{G}_T$ with $[u-h, u+h] \subseteq [0,1]$ such that $m_T'(w) \geq c_T\sqrt{\log T/(Th^3)}$ for all $w \in [u-h, u+h]$. Arguing analogously as in the proof of Proposition 2.3, we get that

$$\begin{aligned}
\widehat{\Psi}_T &\geq \frac{\sqrt{Th}}{\kappa\widehat{\sigma}} \int_0^1 h^{-1} K\Big(\frac{w-u}{h}\Big)\Big(\frac{w-u}{h}\Big) m_T(w)dw + O_p(\sqrt{\log T}) \\
&= \frac{\sqrt{Th^3}}{\kappa\widehat{\sigma}} \int_0^1 h^{-1} K\Big(\frac{w-u}{h}\Big)\Big(\frac{w-u}{h}\Big)^2 m_T'(\xi_{u,w})dw + O_p(\sqrt{\log T}) \\
&\geq \frac{\int_0^1 K(\varphi)\varphi^2 d\varphi}{\kappa\widehat{\sigma}}\, c_T\sqrt{\log T} + O_p(\sqrt{\log T}),
\end{aligned}$$

where $\kappa = \int K^2(\varphi)\varphi^2 d\varphi$ and $\xi_{u,w} \in [u-h, u+h]$ is an intermediate point between u and w. Since $q_T'(\alpha) = O(\sqrt{\log T})$ for any fixed $\alpha \in (0,1)$, this implies that $\mathbb{P}(\widehat{\Psi}_T' \leq q_T'(\alpha)) = o(1)$.

Proof of Theorem 4.1

The proof proceeds analogous to that of Theorem 2.1. In the first step, we show that there exist statistics $\widetilde{\Phi}_{n,T}$ for $T = 1, 2, \ldots$ with the following two properties: (i) $\widetilde{\Phi}_{n,T}$ has the same distribution as $\widehat{\Phi}_{n,T}$ for any T, and (ii)

$$\big|\widetilde{\Phi}_{n,T} - \Phi_{n,T}\big| = o_p\Big(\frac{T^{1/q}}{\sqrt{Th_{\min}}}\Big), \tag{A.13}$$

where $\Phi_{n,T}$ is a Gaussian statistic as defined in Section 4.3. In the second step, we prove that

$$\sup_{x\in\mathbb{R}} \big|\mathbb{P}(\widetilde{\Phi}_{n,T} \leq x) - \mathbb{P}(\Phi_{n,T} \leq x)\big| = o(1). \tag{A.14}$$

To verify (A.13), we follow the arguments for the proof of Proposition A.1, adapting the notation accordingly. The details are given below. The proof of (A.14) is almost identical to that of (A.5) and thus omitted.

We now turn to the proof of (A.13). By Theorem 2.1 and Corollary 2.1 in Berkes et al. (2014), there exist a standard Brownian motion $\mathbb{B}_i$ and a sequence $\{\widetilde{\varepsilon}_{it} : t \in \mathbb{N}\}$ for each i such that the following holds: (i) $\mathbb{B}_i$ and $\{\widetilde{\varepsilon}_{it} : t \in \mathbb{N}\}$ are independent across i, (ii) $[\widetilde{\varepsilon}_{i1}, \ldots, \widetilde{\varepsilon}_{iT}] \stackrel{\mathcal{D}}{=} [\varepsilon_{i1}, \ldots, \varepsilon_{iT}]$ for each i and T, and (iii)

$$\max_{1\leq t\leq T} \Big|\sum_{s=1}^{t} \widetilde{\varepsilon}_{is} - \sigma\mathbb{B}_i(t)\Big| = o\big(T^{1/q}\big) \quad \text{a.s.}$$

for each i, where $\sigma^2 = \sum_{k\in\mathbb{Z}} \mathrm{Cov}(\varepsilon_{i0}, \varepsilon_{ik})$ denotes the long-run error variance. We define

$$\widetilde{\Phi}_T = \max_{1\leq i<j\leq N} \widetilde{\Phi}_{ij,T} \quad \text{with} \quad \widetilde{\Phi}_{ij,T} = \max_{(u,h)\in\mathcal{G}_T} \Big\{\Big|\frac{\widetilde{\phi}_{ij,T}(u,h)}{\sqrt{2}\widetilde{\sigma}}\Big| - \lambda(h)\Big\},$$

where $\widetilde{\phi}_{ij,T}(u,h) = \sum_{t=1}^{T} w_{t,T}(u,h)(\widetilde{\varepsilon}_{it} - \widetilde{\varepsilon}_{jt})$ and $\widetilde{\sigma}$ is the same estimator as $\widehat{\sigma}^\circ$ with $Y_{it}^\circ = m_i(t/T) + \varepsilon_{it}$ replaced by $\widetilde{Y}_{it} = m_i(t/T) + \widetilde{\varepsilon}_{it}$. In addition, we let

$$\Phi_T = \max_{1\leq i<j\leq N} \Phi_{ij,T} \quad \text{with} \quad \Phi_{ij,T} = \max_{(u,h)\in\mathcal{G}_T} \Big\{\Big|\frac{\phi_{ij,T}(u,h)}{\sqrt{2}\sigma}\Big| - \lambda(h)\Big\}$$
$$\Phi_T^* = \max_{1\leq i<j\leq N} \Phi_{ij,T}^* \quad \text{with} \quad \Phi_{ij,T}^* = \max_{(u,h)\in\mathcal{G}_T} \Big\{\Big|\frac{\phi_{ij,T}(u,h)}{\sqrt{2}\widetilde{\sigma}}\Big| - \lambda(h)\Big\},$$

where $\phi_T(u,h) = \sum_{t=1}^{T} w_{t,T}(u,h)\sigma(Z_{it} - Z_{jt})$ and $Z_{it} = \mathbb{B}_i(t) - \mathbb{B}_i(t-1)$. With this notation at hand, we can follow the steps of the proof for Proposition A.1 to arrive at (A.13).

Proof of Proposition 4.3

Consider the event

$$B_{n,T} = \Big\{ \max_{1\le\ell\le N} \max_{i,j\in G_\ell} \widehat{\Psi}_{ij,T} \le q_{n,T}(\alpha) \;\text{ and }\; \min_{1\le\ell<\ell'\le N} \min_{\substack{i\in G_\ell \\ j\in G_{\ell'}}} \widehat{\Psi}_{ij,T} > q_{n,T}(\alpha) \Big\}.$$

The term $\max_{1\le\ell\le N} \max_{i,j\in G_\ell} \widehat{\Psi}_{ij,T}$ is the largest multiscale distance between two time series i and j from the same group, whereas $\min_{1\le\ell<\ell'\le N} \min_{i\in G_\ell, j\in G_{\ell'}} \widehat{\Psi}_{ij,T}$ is the smallest multiscale distance between two time series from two different groups. On the event $B_{n,T}$, it obviously holds that

$$\max_{1\le\ell\le N} \max_{i,j\in G_\ell} \widehat{\Psi}_{ij,T} < \min_{1\le\ell<\ell'\le N} \min_{\substack{i\in G_\ell \\ j\in G_{\ell'}}} \widehat{\Psi}_{ij,T}. \tag{A.15}$$

Hence, any two time series from the same class have a smaller distance than any two time series from two different classes. From Theorem 4.1, it follows that

$$\mathbb{P}\Big(\max_{1\le\ell\le N} \max_{i,j\in G_\ell} \widehat{\Psi}_{ij,T} \le q_{n,T}(\alpha) \Big) \ge (1-\alpha) + o(1).$$

Moreover, by part (b) of Proposition 4.2, we obtain that

$$\mathbb{P}\Big(\min_{1\le\ell<\ell'\le N} \min_{\substack{i\in G_\ell \\ j\in G_{\ell'}}} \widehat{\Psi}_{ij,T} \le q_{n,T}(\alpha) \Big) = o(1).$$

Taken together, these two statements imply that

$$\mathbb{P}\big(B_{n,T}\big) \ge (1-\alpha) + o(1). \tag{A.16}$$

In what follows, we show that on the event $B_{n,T}$, (i) $\{\widehat{G}_1^{[n-N]}, \ldots, \widehat{G}_N^{[n-N]}\} = \{G_1, \ldots$ $\ldots, G_N\}$ and (ii) $\widehat{N} = N$. From (i), (ii) and (A.16), the statements of Proposition 4.3 easily follow.

Proof of (i). Suppose we are on the event $B_{n,T}$. The proof proceeds by induction on the iteration steps r of the HAC algorithm.

Base case ($r = 0$): In the first iteration step, the HAC algorithm merges two singleton clusters $\widehat{G}_i^{[0]} = \{i\}$ and $\widehat{G}_j^{[0]} = \{j\}$ with i and j belonging to the same group G_k. This is a direct consequence of (A.15). The algorithm thus produces a partition $\{\widehat{G}_1^{[1]}, \ldots, \widehat{G}_{n-1}^{[1]}\}$ whose elements $\widehat{G}_\ell^{[1]}$ all have the following property: $\widehat{G}_\ell^{[1]} \subseteq G_k$ for some k, that is, the clusters $\widehat{G}_\ell^{[1]}$ contain elements from only one group.

Induction step ($r \curvearrowright r+1$): Now suppose we are in the r-th iteration step for some $r < n-N$. Assume that the partition $\{\widehat{G}_1^{[r]}, \ldots, \widehat{G}_{n-r}^{[r]}\}$ is such that for any ℓ, $\widehat{G}_\ell^{[r]} \subseteq G_k$

for some k. Because of (A.15), the dissimilarity $\widehat{\Delta}(\widehat{G}_\ell^{[r]}, \widehat{G}_{\ell'}^{[r]})$ gets minimal for two groups $\widehat{G}_\ell^{[r]}$ and $\widehat{G}_{\ell'}^{[r]}$ with the property that $\widehat{G}_\ell^{[r]} \cup \widehat{G}_{\ell'}^{[r]} \subseteq G_k$ for some k. Hence, the HAC algorithm produces a partition $\{\widehat{G}_1^{[r+1]}, \ldots, \widehat{G}_{n-(r+1)}^{[r+1]}\}$ whose elements $\widehat{G}_\ell^{[r+1]}$ are all such that $\widehat{G}_\ell^{[r+1]} \subseteq G_k$ for some k.

The above induction argument shows the following: For any $r \leq n - N$, the partition $\{\widehat{G}_1^{[r]}, \ldots, \widehat{G}_{n-r}^{[r]}\}$ consists of clusters $\widehat{G}_\ell^{[r]}$ which all have the property that $\widehat{G}_\ell^{[r]} \subseteq G_k$ for some k. This in particular holds for the partition $\{\widehat{G}_1^{[n-N]}, \ldots, \widehat{G}_N^{[n-N]}\}$, which in turn implies that $\{\widehat{G}_1^{[n-N]}, \ldots, \widehat{G}_N^{[n-N]}\} = \{G_1, \ldots, G_N\}$. □

Proof of (ii). First consider any partition $\{\widehat{G}_1^{[n-r]}, \ldots, \widehat{G}_r^{[n-r]}\}$ with $r < N$ elements. Such a partition must contain at least one element $\widehat{G}_\ell^{[n-r]}$ with the following property: $\widehat{G}_\ell^{[n-r]} \cap G_k \neq \emptyset$ and $\widehat{G}_\ell^{[n-r]} \cap G_{k'} \neq \emptyset$ for some $k \neq k'$. On the event $B_{n,T}$, it obviously holds that $\widehat{\Delta}(S) > q_{n,T}(\alpha)$ for any S with the property that $S \cap G_k \neq \emptyset$ and $S \cap G_{k'} \neq \emptyset$ for some $k \neq k'$. Hence, we can infer that on the event $B_{n,T}$, $\max_{1 \leq \ell \leq r} \widehat{\Delta}(\widehat{G}_\ell^{[n-r]}) > q_{n,T}(\alpha)$ for any $r < N$.
Next consider the partition $\{\widehat{G}_1^{[n-r]}, \ldots, \widehat{G}_r^{[n-r]}\}$ with $r = N$ and suppose we are on the event $B_{n,T}$. From (i), we already know that $\{\widehat{G}_1^{[n-N]}, \ldots, \widehat{G}_N^{[n-N]}\} = \{G_1, \ldots, G_N\}$. Moreover, it is easy to see that $\widehat{\Delta}(G_\ell) \leq q_{n,T}(\alpha)$ for any ℓ. Hence, we obtain that $\max_{1 \leq \ell \leq N} \widehat{\Delta}(\widehat{G}_\ell^{[n-N]}) = \max_{1 \leq \ell \leq N} \widehat{\Delta}(G_\ell) \leq q_{n,T}(\alpha)$.
Putting everything together, we can conclude that on the event $B_{n,T}$,

$$\min\Big\{r = 1, 2, \ldots \,\Big|\, \max_{1 \leq \ell \leq r} \widehat{\Delta}\big(\widehat{G}_\ell^{[n-r]}\big) \leq q_{n,T}(\alpha)\Big\} = N,$$

that is, $\widehat{N} = N$. □

Proof of Proposition 4.4

For simplicity of notation, suppose that $\alpha_i = 0$ for all i. This allows us to write

$$\widehat{\Phi}_{n,T} = \max_{1 \leq i < j \leq n} \max_{(u,h) \in \mathcal{G}_T} \Big\{ \Big| \frac{\widehat{\psi}_{ij,T}(u,h) - \mathbb{E}\widehat{\psi}_{ij,T}(u,h)}{\sqrt{2}\widehat{\sigma}} \Big| - \lambda(h) \Big\}.$$

We consider the event

$$D_{n,T} = \Big\{ \widehat{\Phi}_{n,T} \leq q_{n,T}(\alpha) \text{ and } \min_{1 \leq \ell < \ell' \leq N} \min_{\substack{i \in G_\ell \\ j \in G_{\ell'}}} \widehat{\Psi}_{ij,T} > q_{n,T}(\alpha) \Big\},$$

which can be analyzed by the same arguments as those applied to the event $B_{n,T}$ in the proof of Proposition 4.3. In particular, analogous to (A.16) and statements (i) and

(ii) therein, we can show that

$$\mathbb{P}\big(D_{n,T}\big) \geq (1-\alpha) + o(1) \tag{A.17}$$

and

$$D_{n,T} \subseteq \big\{\widehat{N} = N \text{ and } \widehat{G}_\ell = G_\ell \text{ for all } \ell\big\}. \tag{A.18}$$

Moreover, we have that

$$D_{n,T} \subseteq \bigcap_{1\leq \ell < \ell' \leq \widehat{N}} E_{n,T}(\ell, \ell'), \tag{A.19}$$

which is a consequence of the following observation: For all i, j and $(u,h) \in \mathcal{G}_T$ with

$$\Big|\frac{\widehat{\psi}_{ij,T}(u,h) - \mathbb{E}\widehat{\psi}_{ij,T}(u,h)}{\sqrt{2}\widehat{\sigma}}\Big| - \lambda(h) \leq q_{n,T}(\alpha) \quad \text{and} \quad \Big|\frac{\widehat{\psi}_{ij,T}(u,h)}{\sqrt{2}\widehat{\sigma}}\Big| - \lambda(h) > q_T(\alpha),$$

it holds that $\mathbb{E}[\widehat{\psi}_{ij,T}(u,h)] \neq 0$, which in turn implies that $m_i(v) - m_j(v) \neq 0$ for some $v \in I_{u,h}$. From (A.18) and (A.19), we obtain that

$$D_{n,T} \subseteq \Big\{ \bigcap_{1\leq \ell < \ell' \leq \widehat{N}} E_{n,T}(\ell, \ell') \Big\} \cap \big\{\widehat{N} = N \text{ and } \widehat{G}_\ell = G_\ell \text{ for all } \ell\big\} = E_{n,T}.$$

This together with (A.17) implies that $\mathbb{P}(E_{n,T}) \geq (1-\alpha) + o(1)$, thus completing the proof.

Proof of Proposition A.3

The proof makes use of the following three lemmas, which correspond to Lemmas 5–7 in Chernozhukov et al. (2015).

Lemma A.5. *Let $(W_1, \ldots, W_p)^\top$ be a (not necessarily centred) Gaussian random vector in $\mathbb{R}^p$ with $\mathrm{Var}(W_j) = 1$ for all $1 \leq j \leq p$. Suppose that $Corr(W_j, W_k) < 1$ whenever $j \neq k$. Then the distribution of $\max_{1\leq j\leq p} W_j$ is absolutely continuous with respect to Lebesgue measure and a version of the density is given by*

$$f(x) = f_0(x) \sum_{j=1}^{p} e^{\mathbb{E}[W_j]x - \mathbb{E}[W_j]^2/2}\, \mathbb{P}\big(W_k \leq x \text{ for all } k \neq j \,\big|\, W_j = x\big).$$

Lemma A.6. *Let $(W_0, W_1, \ldots, W_p)^\top$ be a (not necessarily centred) Gaussian random vector in $\mathbb{R}^p$ with $\mathrm{Var}(W_j) = 1$ for all $1 \leq j \leq p$. Suppose that $\mathbb{E}[W_0] \geq 0$. Then the map*

$$x \mapsto e^{\mathbb{E}[W_0]x - \mathbb{E}[W_0]^2/2}\, \mathbb{P}\big(W_j \leq x \text{ for } 1 \leq j \leq p \,\big|\, W_0 = x\big)$$

is non-decreasing on $\mathbb{R}$.

Lemma A.7. *Let $(X_1, \dots, X_p)^\top$ be a centred Gaussian random vector in $\mathbb{R}^p$ with $\max_{1\leq j\leq p} \mathbb{E}[X_j^2] \leq \sigma^2$ for some $\sigma^2 > 0$. Then for any $r > 0$,*

$$\mathbb{P}\Big(\max_{1\leq j\leq p} X_j \geq \mathbb{E}\Big[\max_{1\leq j\leq p} X_j \Big] + r \Big) \leq e^{-r^2/(2\sigma^2)}.$$

The proof of Lemmas A.5 and A.6 can be found in Chernozhukov et al. (2015). Lemma A.7 is a standard result on Gaussian concentration whose proof is given e.g. in Ledoux (2001); see in particular Theorem 7.1 therein. We now closely follow the arguments for the proof of Theorem 3 in Chernozhukov et al. (2015). The proof splits up into three steps.

Step 1. To start with, we show that the analysis can be restricted to the unit variance case. To see this, pick any $x \geq 0$ and set

$$W_j = \frac{X_j - x}{\sigma_j} + \frac{\overline{\mu} + x}{\underline{\sigma}}.$$

By construction, $\mathbb{E}[W_j] \geq 0$ and $\mathrm{Var}(W_j) = 1$. Defining $Z = \max_{1\leq j\leq p} W_j$, it holds that

$$\begin{aligned}
\mathbb{P}\Big(\Big| \max_{1\leq j\leq p} X_j - x \Big| \leq \delta \Big) &\leq \mathbb{P}\Big(\Big| \max_{1\leq j\leq p} \frac{X_j - x}{\sigma_j} \Big| \leq \frac{\delta}{\underline{\sigma}} \Big) \\
&\leq \sup_{y\in\mathbb{R}} \mathbb{P}\Big(\Big| \max_{1\leq j\leq p} \frac{X_j - x}{\sigma_j} + \frac{\overline{\mu} + x}{\underline{\sigma}} - y \Big| \leq \frac{\delta}{\underline{\sigma}} \Big) \\
&= \sup_{y\in\mathbb{R}} \mathbb{P}\Big(|Z - y| \leq \frac{\delta}{\underline{\sigma}} \Big).
\end{aligned}$$

Step 2. We next bound the density of Z. Without loss of generality, we assume that $\mathrm{Corr}(W_j, W_k) < 1$ for $k \neq j$. The marginal distribution of W_j is $N(\nu_j, 1)$ with $\nu_j = \mathbb{E}[W_j] = (\mu_j/\sigma_j + \overline{\mu}/\underline{\sigma}) + (x/\underline{\sigma} - x/\sigma_j) \geq 0$. Hence, by Lemmas A.5 and A.6, the random variable Z has a density of the form

$$f_p(z) = f_0(z) G_p(z), \tag{A.20}$$

where the map $z \mapsto G_p(z)$ is non-decreasing. Define $\overline{Z} = \max_{1\leq j\leq p}(W_j - \mathbb{E}[W_j])$ and set $\overline{z} = 2\overline{\mu}/\underline{\sigma} + x(1/\underline{\sigma} - 1/\overline{\sigma})$ such that $\mathbb{E}[W_j] \leq \overline{z}$ for any $1 \leq j \leq p$. With these definitions at hand, we obtain that

$$\begin{aligned}
\int_z^\infty f_0(u)du\, G_p(z) \leq \int_z^\infty f_0(u) G_p(u) du &= \mathbb{P}(Z > z) \\
&\leq P(\overline{Z} > z - \overline{z}) \leq \exp\Big(- \frac{(z - \overline{z} - \mathbb{E}[\overline{Z}])_+^2}{2} \Big),
\end{aligned}$$

where the last inequality follows from Lemma A.7. Since $W_j - \mathbb{E}[W_j] = (X_j - \mu_j)/\sigma_j$,

it holds that

$$\mathbb{E}[\overline{Z}] = \mathbb{E}\Big[\max_{1\le j\le p}\Big\{\frac{X_j-\mu_j}{\sigma_j}\Big\}\Big] =: a_p.$$

Hence, for every $z\in\mathbb{R}$,

$$G_p(z) \le \frac{1}{1-F_0(z)}\exp\Big(-\frac{(z-\overline{z}-a_p)_+^2}{2}\Big). \tag{A.21}$$

Mill's inequality states that for $z>0$,

$$z \le \frac{f_0(z)}{1-F_0(z)} \le z\frac{1+z^2}{z^2}.$$

Since $(1+z^2)/z^2\le 2$ for $z>1$ and $f_0(z)/\{1-F_0(z)\}\le 1.53\le 2$ for $z\in(-\infty,1)$, we can infer that

$$\frac{f_0(z)}{1-F_0(z)} \le 2(z\vee 1) \quad \text{for any } z\in\mathbb{R}.$$

This together with (A.20) and (A.21) yields that

$$f_p(z) \le 2(z\vee 1)\exp\Big(-\frac{(z-\overline{z}-a_p)_+^2}{2}\Big) \quad \text{for any } z\in\mathbb{R}.$$

Step 3. By Step 2, we get that for any $y\in\mathbb{R}$ and $u>0$,

$$\mathbb{P}(|Z-y|\le u) = \int_{y-u}^{y+u} f_p(z)dz \le 2u\max_{z\in[y-u,y+u]} f_p(z) \le 4u(\overline{z}+a_p+1),$$

where the last inequality follows from the fact that the map $z\mapsto ze^{-(z-a)^2/2}$ (with $a>0$) is non-increasing on $[a+1,\infty)$. Combining this bound with Step 1, we further obtain that for any $x\ge 0$ and $\delta>0$,

$$\mathbb{P}\Big(\Big|\max_{1\le j\le p} X_j - x\Big|\le\delta\Big) \le 4\delta\Big\{\frac{2\overline{\mu}}{\underline{\sigma}} + |x|\Big(\frac{1}{\underline{\sigma}}-\frac{1}{\overline{\sigma}}\Big) + a_p + 1\Big\}/\underline{\sigma}. \tag{A.22}$$

This inequality also holds for $x<0$ by an analogous argument, and hence for all $x\in\mathbb{R}$.

Now let $0<\delta\le\underline{\sigma}$ and define $b_p = \mathbb{E}\max_{1\le j\le p}\{X_j-\mu_j\}$. For any $|x|\le\delta+\overline{\mu}+b_p+\overline{\sigma}\sqrt{2\log(\underline{\sigma}/\delta)}$, (A.22) yields that

$$\begin{aligned}\mathbb{P}\Big(\Big|\max_{1\le j\le p} X_j - x\Big|\le\delta\Big) &\le \frac{4\delta}{\underline{\sigma}}\Big\{\overline{\mu}\Big(\frac{3}{\underline{\sigma}}-\frac{1}{\overline{\sigma}}\Big) + a_p + \Big(\frac{1}{\underline{\sigma}}-\frac{1}{\overline{\sigma}}\Big)b_p \\ &\qquad + \Big(\frac{\overline{\sigma}}{\underline{\sigma}}-1\Big)\sqrt{2\log\Big(\frac{\underline{\sigma}}{\delta}\Big)} + 2 - \frac{\underline{\sigma}}{\overline{\sigma}}\Big\} \\ &\le C\delta\big\{\overline{\mu}+a_p+b_p+\sqrt{1\vee\log(\underline{\sigma}/\delta)}\big\} \end{aligned}\tag{A.23}$$

with a sufficiently large constant $C > 0$ that depends only on $\underline{\sigma}$ and $\overline{\sigma}$. For $|x| \geq \delta + \overline{\mu} + b_p + \overline{\sigma}\sqrt{2\log(\underline{\sigma}/\delta)}$, we obtain that

$$\mathbb{P}\Big(\Big|\max_{1\leq j\leq p} X_j - x\Big| \leq \delta\Big) \leq \frac{\delta}{\underline{\sigma}}, \tag{A.24}$$

which can be seen as follows: If $x > \delta + \overline{\mu}$, then $|\max_j X_j - x| \leq \delta$ implies that $|x| - \delta \leq \max_j X_j \leq \max_j\{X_j - \mu_j\} + \overline{\mu}$ and thus $\max_j\{X_j - \mu_j\} \geq |x| - \delta - \overline{\mu}$. It thus holds that

$$\mathbb{P}\Big(\Big|\max_{1\leq j\leq p} X_j - x\Big| \leq \delta\Big) \leq \mathbb{P}\Big(\max_{1\leq j\leq p}\big\{X_j - \mu_j\big\} \geq |x| - \delta - \overline{\mu}\Big). \tag{A.25}$$

If $x < -(\delta + \overline{\mu})$, then $|\max_j X_j - x| \leq \delta$ implies that $\max_j\{X_j - \mu_j\} \leq -|x| + \delta + \overline{\mu}$. Hence, in this case,

$$\begin{aligned}\mathbb{P}\Big(\Big|\max_{1\leq j\leq p} X_j - x\Big| \leq \delta\Big) &\leq \mathbb{P}\Big(\max_{1\leq j\leq p}\big\{X_j - \mu_j\big\} \leq -|x| + \delta + \overline{\mu}\Big)\\ &\leq \mathbb{P}\Big(\max_{1\leq j\leq p}\big\{X_j - \mu_j\big\} \geq |x| - \delta - \overline{\mu}\Big),\end{aligned} \tag{A.26}$$

where the last inequality follows from the fact that for centred Gaussian random variables Z_j and $z > 0$, $\mathbb{P}(\max_j Z_j \leq -z) \leq \mathbb{P}(Z_1 \leq -z) = P(Z_1 \geq z) \leq \mathbb{P}(\max_j Z_j \geq z)$. With (A.25) and (A.26), we obtain that for any $|x| \geq \delta + \overline{\mu} + b_p + \overline{\sigma}\sqrt{2\log(\underline{\sigma}/\delta)}$,

$$\begin{aligned}&\mathbb{P}\Big(\Big|\max_{1\leq j\leq p} X_j - x\Big| \leq \delta\Big) \leq \mathbb{P}\Big(\max_{1\leq j\leq p}\big\{X_j - \mu_j\big\} \geq |x| - \delta - \overline{\mu}\Big)\\ &\quad\leq \mathbb{P}\Big(\max_{1\leq j\leq p}\big\{X_j - \mu_j\big\} \geq \mathbb{E}\Big[\max_{1\leq j\leq p}\big\{X_j - \mu_j\big\}\Big] + \overline{\sigma}\sqrt{2\log(\underline{\sigma}/\delta)}\Big) \leq \frac{\delta}{\underline{\sigma}},\end{aligned}$$

the last inequality following from Lemma A.7. To sum up, we have established that for any $0 < \delta \leq \underline{\sigma}$ and any $x \in \mathbb{R}$,

$$\mathbb{P}\Big(\Big|\max_{1\leq j\leq p} X_j - x\Big| \leq \delta\Big) \leq C\delta\big\{\overline{\mu} + a_p + b_p + \sqrt{1 \vee \log(\underline{\sigma}/\delta)}\big\} \tag{A.27}$$

with some constant $C > 0$ that does only depend on $\underline{\sigma}$ and $\overline{\sigma}$. For $\delta > \underline{\sigma}$, (A.27) trivially follows upon setting $C \geq 1/\underline{\sigma}$. This completes the proof.

References

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

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.

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.

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.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning*. New York, Springer.

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

Ledoux, M. (2001). *Concentration of Measure Phenomenon*. American Mathematical Society.

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

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.

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.