Multiscale Inference for Nonparametric Time Trends

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 $\varepsilon_t = G(\ldots, e_{t-1}, e_t, e_{t+1}, \ldots)$, where e_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 $d_{t,q} = \|\varepsilon_t - \varepsilon_t^\circ\|_q$, where $\varepsilon_t^\circ = G(\ldots, e_1, e_0', e_1, \ldots, e_{t-1}, e_t)$ with $\{e_t'\}$ being an i.i.d. copy of $\{e_t\}$. In particular, we assume the following:

(C3) It holds that

$$\sum_{t=1}^{\infty} t\|\varepsilon_t - \varepsilon_t^\circ\|_q < \infty.$$

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 $\varepsilon_t = \sum_{i=0}^{\infty} c_i e_{t-i}$, where c_i are absolutely summable coefficients and e_t are i.i.d. innovations with $\mathbb{E}[e_t] = 0$ and $\|e_t\|_q < \infty$. Obviously, (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, let $\{\varepsilon_t\}$ be an ARMA process of the form $\varepsilon_t + \sum_{i=1}^{p} a_i \varepsilon_{t-i} = e_t + \sum_{j=1}^{r} b_j e_{t-j}$, where $a_1, \ldots, a_p$ and $b_1, \ldots, b_r$ are real-valued parameters and 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 is stationary and causal. Specifically, it has the representation $\varepsilon_t = \sum_{i=0}^{\infty} c_i e_{t-i}$ with $|c_i| = O(\rho^i)$ for some $\rho \in (0,1)$, implying that (C1)–(C3) are fulfilled. As shown in Wu and Shao

(2004), the 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 $\mathcal{Y}_i = \{Y_{it} : 1 \le t \le T\}$ of length T for $1 \le i \le N$. 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 $1 \le t \le T$, where m_i is an unknown nonparametric function defined on $[0,1]$, α_i is a (deterministic or random) intercept term and $\mathcal{E}_i = \{\varepsilon_{it} : 1 \le t \le T\}$ is a zero-mean stationary error process. For identification purposes we normalize the functions m_i such that $\int_0^1 m_i(u)du = 1$ for all $1 \le i \le N$. The conditions on the error processes $\mathcal{E}_i$ can be summarized as follows: The processes $\mathcal{E}_i$ are independent across i and each process $\mathcal{E}_i$ satisfies the conditions (C1)–(C3). We thus work with essentially the same error structure as in our analysis in Sections 2 and 3.

2 The multiscale test procedure

In this section, we introduce our multiscale test method and the underlying theory for the simple hypothesis $H_0 : m = 0$ in the 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 the model (1.1), we consider the kernel averages

$$\widehat{\psi}_T(u,h) = \sum_{t=1}^{T} w_{t,T}(u,h)Y_t, \tag{2.1}$$

where $w_{t,T}(u,h)$ is a kernel weight with $u \in [0,1]$ and h being the bandwidth parameter. We in particular set

$$w_{t,T}(u,h) = \frac{1}{\|K\|_{u,h,T}} K\Big(\frac{u - t/T}{h}\Big) \quad \text{with} \quad \|K\|_{u,h,T} = \Big\{\sum_{t=1}^{T} K^2\Big(\frac{u - t/T}{h}\Big)\Big\}^{1/2},$$

where K is a non-negative kernel function which is symmetric about zero, integrates to one and has compact support $[-1,1]$. 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. Defining the long-run error variance by σ^2, we have $\mathrm{Var}(\widehat{\psi}_T(u,h)) = \sigma^2 + o(1)$ for any fixed location u and any bandwidth h with $h \to 0$ and $Th \to \infty$, meaning that the statistics $\widehat{\psi}_T(u,h)$ should have approximately the same variance across u and h for sufficiently large sample sizes T. In what follows, we mainly 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. For the time being, we suppose that $\widehat{\sigma}^2$ is an estimator with reasonable theoretical properties. In particular, we assume that $\widehat{\sigma}^2 = \sigma^2 + O_p(1/\sqrt{T})$.
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 takes into account all points $(u,h) \in \mathcal{G}_T$, where $\mathcal{G}_T$ is some subset of

$$\mathcal{G} = \big\{(u,h) : [u-h, u+h] \subseteq [0,1] \text{ with } u \in [0,1] \text{ and } h \in [h_{\min}, h_{\max}]\big\}$$

with $h_{\min}$ and $h_{\max}$ denoting some minimal and maximal bandwidth value respectively. In order our theory to work, we require the following conditions to hold:

(C4) $|\mathcal{G}_T| = O(T^\gamma)$ for some arbitrarily large but fixed constant $\gamma > 0$, where $|\mathcal{G}_T|$ denotes the cardinality of $\mathcal{G}_T$.

(C5) $h_{\min} \gg T^{(2-q)/q}$ that is, $h_{\min}/T^{(2-q)/q} \to \infty$ with $q > 4$ defined in (C2) and $h_{\max} < 1/2$.

According to (C4), the number of points (u,h) in $\mathcal{G}_T$ should not grow faster than T^γ for some arbitrarily large but fixed $\gamma > 0$. 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

$$\begin{aligned}\mathcal{G}_T = \big\{(u,h) : [u-h, u+h] \subseteq [0,1] \text{ with } u = t/T \text{ for some } 1 \le t \le T \\ \text{and } h \in [h_{\min}, h_{\max}] \text{ with } h = t/T \text{ for some } 1 \le t \le T\big\}\end{aligned}$$

which contains more than enough points (u,h) for most practical applications. Condition (C5) imposes restrictions on the minimal and maximal bandwidths $h_{\min}$ and $h_{\max}$ used in our multiscale approach. However, 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 $h_{\min}$ converges 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.
Following the approach in Dümbgen and Spokoiny (2001), we define our multiscale

statistic 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 $\lambda(h) = \sqrt{2\log\{1/(2h)\}}$. As suggested there, we thus do 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 the traditional approach; we rather subtract the additive correction term $\lambda(h)$ from the statistics 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

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

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 bandwidth values 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.

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

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

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

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

with

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

According to the following theorem, for any given $\alpha \in (0,1)$, the $(1-\alpha)$-quantile of the statistic $\widehat{\Phi}_T$ is approximately equal to the (known) quantile $q_T(\alpha)$ of Φ_T^* defined in Section 2.2.

Theorem 2.1. *Let (C1)–(C5) be fulfilled and suppose that the kernel K is Lipschitz continuous and has compact support $[-1,1]$. 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

$$\big|\widetilde{\Phi}_T - \Phi_T^*\big| = o_p\Big(\frac{T^{1/q}}{\sqrt{Th_{\min}}}\Big), \tag{2.2}$$

where the Gaussian statistic Φ_T^* is defined in Section 2.2. We thus replace the test 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

$$\sup_{x\in\mathbb{R}} \big|\mathbb{P}(\widetilde{\Phi}_T \le x) - \mathbb{P}(\Phi_T^* \le x)\big| = o(1), \tag{2.3}$$

which implies that that for any given $\alpha \in (0,1)$, the $(1-\alpha)$-quantile of the statistic $\widetilde{\Phi}_T$ can be approximated by the known quantile $q_T(\alpha)$ of the Gaussian statistic Φ_T^*. The main tools for verifying (2.3) are anti-concentration results for Gaussian random vectors as derived in Chernozhukov et al. (2015). Combining (2.2) and (2.3), we finally arrive at the statement of Theorem 2.1.

With the help of Theorem 2.1, we can now 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.

Corollary 2.2. *Let the conditions of Theorem 2.1 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. According to it, the test has asymptotic power 1 against fixed alternatives and is thus consistent.

Corollary 2.3. *Let the conditions of Theorem 2.1 be satisfied and let m be any fixed function with $m \neq 0$. Then*

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

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 $|\widehat{\psi}_T(u,h)| - \lambda(h)$ 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.

Corollary 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 Corollary 2.4, our test procedure allows us to make uniform confidence statements of the following form: With (asymptotic) probability at least $(1-\alpha)$, the

null hypothesis $H_0 : m = 0$ is violated on all the 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 Corollary 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 Π_T^* which is constructed as follows: As in ??, 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 Π_T^* be the collection of all minimal intervals in Π_T and define the event

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

It is easily seen that $E_T = E_T^*$. Hence, by Corollary 2.4, it holds that

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

This suggests to plot the minimal intervals in Π_T^* 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 this section we adapt the multiscale test method developed in Section 2.2 to test more interesting hypothesis H_0
Consider the time trend model

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

for $t = 1, \ldots, T$ with $\mathbb{E}[\varepsilon_t] = 0$. We want to test the null hypothesis

$$H_0 : m \text{ is constant.}$$

As in Section subsec-method-test, this can be achieved with the help of the following multiscale test statistic:

$$\mathcal{M}_T = \max_{(u,h)\in\mathcal{G}_T} \Big\{|\Psi_T(u,h)| - \lambda(h)\},$$

where

$$\Psi_T(u,h) = \sum_{t=1}^{T} w_{t,T}(u,h)\varepsilon_t$$

with

$$w_{t,T}(u,h) = \frac{1}{\hat{\sigma}\sqrt{Th}}\, K'\Big(\frac{u - t/T}{h}\Big),$$

where $\hat{\sigma}^2$ is a suitable estimator of the long-run error variance σ^2.

4 Testing for equality of time trends

Suppose we observe n different time series $\mathcal{Y}_i = \{Y_{it} : 1 \le t \le T\}$ for $1 \le i \le n$. The i-th time series is assumed to follow the time trend model

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

for $1 \le t \le T$, where $\mathbb{E}[\varepsilon_{it}] = 0$ and α_i are (deterministic or random) intercept terms. (We here set $\alpha_i = 0$ for simplicity and incorporate the intercepts α_i later into the procedure.) We want to test the null hypothesis that

$$H_0 : m_1 = m_2 = \ldots = m_n.$$

To do so, we modify the baseline procedure from Section ??: For any pair of time series i and j, we define

$$\mathcal{M}_{ij,T} = \max_{L\in\mathcal{L}} \max_{\substack{1\le t\le t'\le T \\ |t'-t+1|=L}} \Big\{ |\Delta_{ij}(t,t')| - \sqrt{2\log\left(\tfrac{T}{L}\right)} \Big\},$$

where

$$\Delta_{ij}(t,t') = \frac{1}{\sqrt{|t'-t+1|}} \sum_{s=t}^{t'} V_s \quad \text{with} \quad V_{ij,s} = \frac{(Y_{is} - \mathbb{E}[Y_{is}]) - (Y_{js} - \mathbb{E}[Y_{js}])}{2\sigma}$$

for $t' \ge t$. The test statistic $\mathcal{M}_T$ is then defined as

$$\mathcal{M}_T = \max_{1\le i<j\le n} \mathcal{M}_{ij,T}.$$

5 Estimation of the long-run error variance

For estimating the long-run error covariance $\sigma^2 = \sum_{k\in\mathbb{Z}} \mathbb{E}[\varepsilon_0\varepsilon_k]$ various consistent estimators can be used. Here we employ the difference-based method proposed by **?** which has the advantage that it does not depend on a bandwidth parameter. The exact procedure of constructing $\hat{\sigma}^2$ is described below.

As in the previous sections, we allow for dependency structure in the errors. Specifically, we assume that the errors $\epsilon_1, \dots, \epsilon_T$ have the following linear structure:

$$\epsilon_t = \sum_{s=1}^{p} \phi_s \epsilon_{t-s} + e_t$$

for some $p \le 1$, where $\{e_t\}_{-\infty}^{+\infty}$ are i.i.d. with $\mathbb{E}[e_t] = 0$ and $\mathrm{Var}[e_t] = \sigma_e^2 < \infty$. Furthermore, we assume that the process $\{\epsilon_t\}$ is causal, that is the constants $\phi_1, \dots, \phi_p$ are such that the equation $1 - \sum_j \phi_j z^j = 0$ has no complex roots inside the unit circle. Under these assumptions we can construct a $\sqrt{T}$-consistent estimator $\hat{\sigma}^2$ in three following steps.

Step 1. First we focus our attention on the error covariance at lag s: $\gamma_s = cov(\epsilon_t, \epsilon_{t-s})$. Here we exploit pairwise differences of the Y_ts: for a given $s \in \mathbb{Z}$ define the difference operator $D_s : (D_s Y)_t = Y_t - Y_{t-s}$. Then the estimators are constructed as follows:

$$\hat{\gamma}(0) = \frac{1}{m_2 - m_1 + 1} \sum_{m=m_1}^{m_2} \frac{1}{2(T-m)} \sum_{t=m+1}^{T} \{(D_m Y)_t\}^2,$$

$$\hat{\gamma}(s) = \hat{\gamma}(0) - \frac{1}{2(T-s)} \sum_{t=s+1}^{T} \{(D_s Y)_t\}^2 \text{ for } s \ge 1,$$

$$\hat{\gamma}(s) = \hat{\gamma}(-s) \text{ for } s \leq -1,$$

where $m_1 \leq m_2$ are subsidiary smoothing parameters.

Step 2. Now we can estimate coefficients of the eror model ϕ_s by using the Yule-Walker equations:

$$\gamma(s) = \sum_{r \geq 1} \phi_r \gamma(s - r).$$

Then denoting A the $p \times p$ matrix of error covariance with elements $a_{s_1,s_2} = \gamma(s_1 - s_2)$ and replacing γ by $\hat{\gamma}$ in the definition of A to obtain an estimator $\hat{A}$, straightforwardly we have estimators for the coefficients of the model:

$$(\hat{\phi}_1, \ldots, \hat{\phi}_p)^T = \hat{A}^{-1}(\hat{\gamma}(1), \ldots, \hat{\gamma}(p))^T$$

Step 3. The estimators $\hat{\gamma}(s)$ from the previous step do not exploit the structure of the process ϵ_t which can lead to unreasonably high variability of $\hat{\gamma}(s)$ for large s. For this reason we can modify the approach in Step 1 so that it extracts more infromation from the dependency structure between the errors ϵ_t.

Once the estimators $\hat{\phi}_1, \ldots, \hat{\phi}_p$ are constructed, define $\hat{\psi}_1, \hat{\psi}_1, \ldots$ by

$$1 + \sum_{j \geq 1} \hat{\psi}_j z^j = \Big(1 - \sum_{j=1}^{p} \hat{\phi}_j z^j\Big)^{-1}$$

and let $\hat{\psi}_0 = 1$. Furthermore, define $\bar{\gamma}(0), \bar{\gamma}(1), \ldots$ by

$$\bar{\gamma}(s) = \sum_{k,l \geq 0, k+l=s} \sum \hat{\psi}(k)\hat{\psi}(l).$$

Then $\hat{\sigma}_e^2 = \hat{\gamma}(0)\bar{\gamma}(0)$ would be a suitable estimator of $\sigma_e^2 = \mathrm{Var}[e_t]$. And if we define $\tilde{\gamma}(s) = \hat{\sigma}_e^2 \bar{\gamma}(s)$ for $s \geq 1$, we will obtain an estimator for $\sigma^2 = \sum_{k \in \mathbb{Z}} \mathbb{E}[\epsilon_0 \epsilon_k] = \gamma(0) + 2\sum_{s \geq 1} \gamma(s)$ using the following formula:

$$\hat{\sigma} = \hat{\sigma}_e^2 \Big(1 - \sum_{1 \leq j \leq p} \hat{\psi}_j\Big)^{-2}.$$

It may be proved that if the error distribution has finite fourth moment, if smoothing parameters are such that $m_1 \leq m_2, m_1/\log(T) \to \infty$ and $m_2 = O(\sqrt{T})$ and under mild smoothness conditions on $m(x)$ we have

$$\max_{0 \leq s \leq T} |\hat{\gamma}(s) - \gamma(s)| = O_p(\frac{1}{\sqrt{T}}) \text{ and}$$
$$\max_{0 \leq j \leq p} |\hat{\phi}(j) - \phi(j)| = O_p(\frac{1}{\sqrt{T}})$$

$$\hat{\sigma}^2 - \sigma^2 = O_p(\frac{1}{\sqrt{T}})$$

Appendix

In this appendix, 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 Gaussian distribution, respectively.

Proof of Theorem 2.1

As already outlined in Section 2.3, the proof splits up into two main steps. In the first, we use strong approximation theory to show 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)*

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

where Φ_T^* is a Gaussian statistic as defined in Section 2.2.

According to this result, we can replace the statistic $\widehat{\Phi}_T$ by an identically distributed version $\widetilde{\Phi}_T$ which is close to the Gaussian statistic Φ_T^*. We defer the proof of Proposition A.1 until the arguments for Theorem 2.1 are completed. In the second main step of the proof, we show that

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

which immediately implies the statement of Theorem 2.1. For the proof of (A.1), we use the following simple lemma:

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

$$\sup_{x\in\mathbb{R}} \mathbb{P}(|V_T - x| \leq \delta_T) = o(1), \tag{A.2}$$

then

$$\sup_{x\in\mathbb{R}} \big|\mathbb{P}(V_T \leq x) - \mathbb{P}(W_T \leq x)\big| = o(1). \tag{A.3}$$

The statement of this lemma 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.2), then the distribution of W_T can be approximated by that of V_T in the sense of (A.3).

Proof of Lemma A.2. It holds that

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

□

We now apply Lemma A.2 with $V_T = \Phi_T^*$, $W_T = \widetilde{\Phi}_T$ and $\delta_T = T^{1/q}/\sqrt{Th_{\min}}$: Using Lévy's modulus of continuity, it can be shown that $\Phi_T^* = O_p(1)$. Moreover, from Proposition A.1, we already know that $\widetilde{\Phi}_T - \Phi_T^* = o_p(\delta_T)$. Finally, with the help of recent anti-concentration bounds for Gaussian random vectors, we can verify the following proposition.

Proposition A.3. *Under the conditions of Theorem 2.1, it holds that*

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

The proof of Proposition A.3 is given below. According to it, the Gaussian multiscale statistic Φ_T^* does not concentrate too strongly in regions of the form $[x-\delta_T, x+\delta_T]$. Putting everything together, we are now in a position to apply Lemma A.2, which in turn yields (A.1). This completes the proof of Theorem 2.1.

Proof of Proposition A.1 – strong approximation theory

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 exists a standard Brownian motion $\mathbb{B}$ and a sequence $\{\widetilde{\varepsilon}_t : 1 \le t \le T\}$ with $[\widetilde{\varepsilon}_1, \ldots, \widetilde{\varepsilon}_T] \stackrel{\mathcal{D}}{=} [\varepsilon_1, \ldots, \varepsilon_T]$ such that

$$\max_{1\le t\le T} \Big|\sum_{s=1}^{t} \widetilde{\varepsilon}_s - \sigma\mathbb{B}(t)\Big| = o\big(T^{1/q}\big) \quad \text{a.s.}, \tag{A.4}$$

where $\sigma^2 = \sum_{k\in\mathbb{Z}} \mathbb{E}[\varepsilon_0\varepsilon_k]$ denotes the long-run error variance. To apply this result, we set

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

with $\widetilde{\phi}_T(u,h) = \sum_{t=1}^T w_{t,T}(u,h)\widetilde{\varepsilon}_t$ as well as

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

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 at hand, we get that

$$\big|\widetilde{\Phi}_T - \Phi_T^*\big| \leq \big|\widetilde{\Phi}_T - \Phi_T^{**}\big| + \big|\Phi_T^{**} - \Phi_T^*\big| = \big|\widetilde{\Phi}_T - \Phi_T^{**}\big| + O_p\Big(\sqrt{\frac{\log T}{T}}\Big), \tag{A.5}$$

where the last equality holds due to the fact that the variables Z_t are independent standard normal and $|\mathcal{G}_T| = O(T^r)$ for some large but fixed constant r. Moreover, straightforward calculations yield that

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

Using summation by parts, we obtain that

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

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.4), we can thus infer that

$$\big|\widetilde{\Phi}_T - \Phi_T^{**}\big| \leq \widehat{\sigma}^{-1} \max_{(u,h)\in\mathcal{G}_T} \big|\widetilde{\phi}_T(u,h) - \phi_T^*(u,h)\big| = o_p\Big(\frac{T^{1/q}}{\sqrt{Th_{\min}}}\Big). \tag{A.6}$$

Plugging (A.6) into (A.5) completes the proof.

Proof of Proposition A.3 – anti-concentration bounds

The main technical tools for proving Proposition A.3 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.4. *Let $(X_1, \ldots, X_p)^\top$ 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 $\overline{\mu} = \max_{1\leq j\leq p} |\mu_j|$ and $a_p = \mathbb{E}[\max_{1\leq j\leq p}(X_j - \mu_j)/\sigma_j]$ as well as $\underline{\sigma} = \min_{1\leq j\leq p} \sigma_j$ and $\overline{\sigma} = \max_{1\leq j\leq p} \sigma_j$. For every $\delta > 0$, it holds that*

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

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

To apply Proposition A.4 to our setting at hand, we introduce the following notation: We write $x_j = (x_{j1}, x_{j2}) = (u, h)$ along with $\{x_1, \ldots, x_p\} = \{x : x \in \mathcal{G}_T\} = \mathcal{G}_T$, where $p := |\mathcal{G}_T| \leq O(T^r)$ for some large but fixed $r > 0$ by our assumptions. Moreover, for $j = 1, \ldots, p$, we set

$$X_{2j-1} = \frac{\phi_T^*(x_{j1}, x_{j2})}{\sigma} - \lambda(x_{j2}) \quad \text{and} \quad X_{2j} = -\frac{\phi_T^*(x_{j1}, x_{j2})}{\sigma} - \lambda(x_{j2}).$$

This notation allows us to write

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

where $(X_1, \ldots, X_{2p})^\top$ 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) $\mathrm{Var}(X_j) = \sigma_j^2$ with $0 < \underline{\sigma} \leq \sigma_j \leq \overline{\sigma} < \infty$ for some fixed constants $\underline{\sigma}$, $\overline{\sigma}$ that are independent of T. Since the variables $(X_j - \mu_j)/\sigma_j$ are standard normal, it holds that $a_p \leq \sqrt{2\log(2p)} \leq C\sqrt{\log T}$. With this notation at hand, we can apply Proposition A.4 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.3.

Proof of Proposition A.4. 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 $(W_0, W_1, \dots, W_p)^\top$ be a (not necessarily centred) Gaussian random vector in $\mathbb{R}^p$ with $\mathrm{Var}(W_j) = 1$ for all $0 \le j \le p$. Suppose that $\mathbb{E}[W_0] \ge 0$. Then the map*

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

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 $\max_{1\le j\le p} \mathbb{E}[X_j^2] \le \sigma^2$ for some $\sigma^2 > 0$. Then for any $r > 0$,*

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

The proofs 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 which 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. Pick any $x \ge 0$ and set

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

By construction, $\mathbb{E}[W_j] \ge 0$ and $\mathrm{Var}(W_j) = 1$. Defining $Z = \max_{1\le j\le p} W_j$, we have

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

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

$$f_p(z) = f_0(z) G_p(z), \tag{A.7}$$

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

definitions at hand, we obtain that

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

where the last inequality is due to 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 $z \in \mathbb{R}$,

$$G_p(z) \le \frac{1}{1 - F_0(z)} \exp\Big(- \frac{(z - \overline{z} - a_p)_+^2}{2}\Big). \tag{A.8}$$

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.7) and (A.8) 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, for any $y \in \mathbb{R}$ and $u > 0$, we have

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

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 get that for any $x \ge 0$ and $\delta > 0$,

$$\mathbb{P}\Big(\Big| \max_{1\le j\le p} X_j - x \Big| \le \delta \Big) \le 4\delta \Big\{ \frac{2\overline{\mu}}{\underline{\sigma}} + |x| \Big(\frac{1}{\underline{\sigma}} - \frac{1}{\overline{\sigma}} \Big) + a_p + 1 \Big\} / \underline{\sigma}. \tag{A.9}$$

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

Now let $0 < \delta \le \underline{\sigma}$. For any $|x| \le \delta + \overline{\mu} + \overline{\sigma}(a_p + \sqrt{2\log(\underline{\sigma}/\delta)})$, (A.9) yields that

$$\begin{aligned}\mathbb{P}\Big(\Big|\max_{1\le j\le p} X_j - x\Big| \le \delta\Big) &\le \frac{4\delta}{\underline{\sigma}}\Big\{\overline{\mu}\Big(\frac{3}{\underline{\sigma}} - \frac{1}{\overline{\sigma}}\Big) + \frac{\overline{\sigma}}{\underline{\sigma}}a_p + \Big(\frac{\overline{\sigma}}{\underline{\sigma}} - 1\Big)\sqrt{2\log\Big(\frac{\underline{\sigma}}{\delta}\Big)} + 2 - \frac{\underline{\sigma}}{\overline{\sigma}}\Big\} \\ &\le C\delta\big\{\overline{\mu} + a_p + \sqrt{1 \vee \log(\underline{\sigma}/\delta)}\big\} \qquad \text{(A.10)}\end{aligned}$$

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

$$\mathbb{P}\Big(\Big|\max_{1\le j\le p} X_j - x\Big| \le \delta\Big) \le \frac{\delta}{\underline{\sigma}}, \qquad \text{(A.11)}$$

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

$$\mathbb{P}\Big(\Big|\max_{1\le j\le p} X_j - x\Big| \le \delta\Big) \le \mathbb{P}\Big(\max_{1\le j\le p}\big\{X_j - \mu_j\big\} \ge |x| - \delta - \overline{\mu}\Big). \qquad \text{(A.12)}$$

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

$$\begin{aligned}\mathbb{P}\Big(\Big|\max_{1\le j\le p} X_j - x\Big| \le \delta\Big) &\le \mathbb{P}\Big(\max_{1\le j\le p}\big\{X_j - \mu_j\big\} \le -|x| + \delta + \overline{\mu}\Big) \\ &\le \mathbb{P}\Big(\max_{1\le j\le p}\big\{X_j - \mu_j\big\} \ge |x| - \delta - \overline{\mu}\Big), \qquad \text{(A.13)}\end{aligned}$$

where the last inequality follows from the fact that for centred Gaussian random variables Z_j and $\forall z > 0$, $\mathbb{P}(\max_j Z_j \le -z) \le \mathbb{P}(Z_1 \le -z) = P(Z_1 \ge z) \le \mathbb{P}(\max_j Z_j \ge z)$. With (A.12) and (A.13), we obtain that for any $|x| \ge \delta + \overline{\mu} + \overline{\sigma}(a_p + \sqrt{2\log(\underline{\sigma}/\delta)})$,

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

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

$$\mathbb{P}\Big(\Big|\max_{1\le j\le p} X_j - x\Big| \le \delta\Big) \le C\delta\big\{\overline{\mu} + a_p + \sqrt{1 \vee \log(\underline{\sigma}/\delta)}\big\} \qquad \text{(A.14)}$$

with some constant $C > 0$ that does only depend on $\underline{\sigma}$ and $\overline{\sigma}$. For $\delta > \underline{\sigma}$, (A.14) trivially follows upon setting $C \ge 1/\underline{\sigma}$. This completes the proof. □

Proof of Corollary 2.4

The statement of Corollary 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.

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

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

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.