The Annals of Statistics
2012, Vol. 40, No. 1, 494–528
DOI: 10.1214/11-AOS963

SIMULTANEOUS CONFIDENCE BANDS FOR YULE–WALKER ESTIMATORS AND ORDER SELECTION

BY MORITZ JIRAK

Graz University of Technology

Let $\{X_k, k \in \mathbb{Z}\}$ be an autoregressive process of order q. Various estimators for the order q and the parameters $\mathbf{\Theta}_q = (\theta_1, \dots, \theta_q)^T$ are known; the order is usually determined with Akaike's criterion or related modifications, whereas Yule–Walker, Burger or maximum likelihood estimators are used for the parameters $\mathbf{\Theta}_q$. In this paper, we establish simultaneous confidence bands for the Yule–Walker estimators $\widehat{\theta}_i$; more precisely, it is shown that the limiting distribution of $\max_{1 \leq i \leq d_n} |\widehat{\theta}_i - \theta_i|$ is the Gumbel-type distribution $e^{-e^{-z}}$, where $q \in \{0, \dots, d_n\}$ and $d_n = \mathcal{O}(n^\delta)$, $\delta > 0$. This allows to modify some of the currently used criteria (AIC, BIC, HQC, SIC), but also yields a new class of consistent estimators for the order q. These estimators seem to have some potential, since they outperform most of the previously mentioned criteria in a small simulation study. In particular, if some of the parameters $\{\theta_i\}_{1 \leq i \leq d_n}$ are zero or close to zero, a significant improvement can be observed. As a byproduct, it is shown that BIC, HQC and SIC are consistent for $q \in \{0, \dots, d_n\}$ where $d_n = \mathcal{O}(n^\delta)$.

1. Introduction. Let $\{X_k\}_{k \in \mathbb{Z}}$ be a qth-order autoregressive process AR(q) with coefficient vector $\mathbf{\Theta}_q \in \mathbb{R}^q$. A considerable literature in the past years dealt with various aspects and problems on AR(q)-processes; see, for instance, [4, 17, 23, 29] and the references therein. More recently, people have moved on to more complicated models such as ARCH [14, 19], GARCH [13] and related models, which again have been extended in many different directions. However, in many applications, AR(q)-processes still form the backbone and are often used as first approximations for further analysis; in particular, many estimation and fitting procedures can be based on preliminary AR$(0q)$ approximations. This includes, for instance, ARMA, ARCH and GARCH models [11, 22, 24]. Thus, AR(q) processes have moved from the spotlight to the backstage area, yet their significance remains unchallenged.

When fitting an AR(q) model, two important questions arise: how to choose the order q, and having done so, which estimators are to be used. Naturally, these two problems can hardly be separated and are often dealt with simultaneously, or at least so in preliminary estimates. An extensive literature has evolved around

Received December 2010; revised September 2011.

MSC2010 subject classifications. Primary 60M10, 62F05; secondary 62F10, 62F12.

Key words and phrases. Autoregressive process, Yule–Walker estimators, extreme value distribution, order selection, AIC.

these two issues. Pioneering contributions in this direction are due to Akaike [1, 2], Mallows [30, 31], Walker [44] and Yule [49]; for more details we refer to [4, 15, 17, 23, 29] and the references there. In order to be able to describe some of the basic results, we recall that an AR(q) process $\{X_k\}_{k\in\mathbb{Z}}$ is defined through the recurrence relation

$$X_k = \theta_1 X_{k-1} + \cdots + \theta_q X_{k-q} + \varepsilon_k, \tag{1.1}$$

where it is often assumed that $\{\varepsilon_k\}_{k\in\mathbb{Z}}$ is a mean-zero i.i.d. sequence. Let $\phi_h = \mathbb{E}(X_k X_{k+h})$, $k, h \in \mathbb{Z}$, be the covariance function. A natural estimate for ϕ_h is the sample covariance $\widehat{\phi}_{n,h} = \frac{1}{n}\sum_{i=h+1}^{n} X_i X_{i-h}$. Depending on the magnitude of h, a different normalization, such as $(n-h)^{-1}$, is sometimes more convenient. Denote with $\mathbf{\Theta}_q = (\theta_1, \ldots, \theta_q)^T$ the parameter vector and put $\mathbf{\Phi}_q = (\phi_1, \ldots, \phi_q)^T$, and let $\mathbf{\Gamma}_q = (\phi_{|i-j|})_{1\le i,j\le q}$ be the $(q \times q)$-dimensional covariance matrix. Then it follows from (1.1) that $\mathbf{\Gamma}_q\mathbf{\Theta}_q = \mathbf{\Phi}_q$; hence a natural idea is to replace the corresponding quantities by estimators $\widehat{\mathbf{\Phi}}_q = (\widehat{\phi}_{n,1}, \ldots, \widehat{\phi}_{n,q})^T$, $\widehat{\mathbf{\Gamma}}_q = (\widehat{\phi}_{n,|i-j|})_{1\le i,j\le q}$, and thus define the estimator $\widehat{\mathbf{\Theta}}_q = (\widehat{\theta}_1, \ldots, \widehat{\theta}_q)^T$ via

$$\widehat{\mathbf{\Gamma}}_q^{-1}\widehat{\mathbf{\Phi}}_q = \widehat{\mathbf{\Theta}}_q \quad \text{and} \quad \widehat{\sigma}^2(q) = \widehat{\phi}_0 - \widehat{\mathbf{\Theta}}_q^T\widehat{\mathbf{\Phi}}_q, \tag{1.2}$$

where $\sigma^2 = \mathbb{E}(\varepsilon_0^2)$. These estimators are commonly referred to as the Yule–Walker estimators, and they have some remarkable properties. For example, if $\{X_k\}_{k\in\mathbb{Z}}$ is causal, then the fitted model

$$X_k = \widehat{\theta}_1 X_{k-1} + \cdots + \widehat{\theta}_p X_{k-q} + \varepsilon_k$$

is still causal; see, for instance, [17] and [34]. Another interesting feature is that even though the Yule–Walker estimators are obtained via moment matching methods, their variance is asymptotically equivalent with those obtained via a maximum likelihood approach. More precisely, for $m \ge q$ it holds that

$$\sqrt{n}(\widehat{\mathbf{\Theta}}_m - \mathbf{\Theta}_m) \xrightarrow{d} \mathcal{N}(0, \sigma^2\mathbf{\Gamma}_m^{-1}), \tag{1.3}$$

where $\mathbf{\Theta}_m = (\theta_1, \ldots, \theta_q, 0, \ldots, 0)^T$; see, for instance, [17]. These asymptotic results form the basis for earlier estimation methods of the order q [37, 43, 45], which focused on a fixed, finite number of possible orders and consist of multiple-testing-procedures, which in practice leads to the difficulty of having a required level. On the other hand, as it was pointed out by Shwarz [39], a direct likelihood approach fails, since it invariably chooses the highest possible dimension. Akaike [1] and Mallows [30, 31], developed a different approach, which is based on a "generalized likelihood function." Shibata [41] investigated the asymptotic distribution and showed that the estimator based on (1.4) is not consistent. This issue was successfully dealt with by Akaike [2] (BIC), Hannan and Quinn [25] (HQC), Parzen [36], Rissanen [38] and Shwarz [39] (SIC), who introduced consistent modifications (Parzen's CAT-criterion is conceptually different). For more recent advances and generalizations, see, for instance, Barron et al. [6], Foster and George [20], Shao [40] and the detailed review on model selection given by Leeb and Pötscher [28]. A particularly interesting direction addresses AR(∞) approximations; recent contributions are due to Bickel and Yel [12] and Ing and Wei [26, 27]. Here and now, we will content ourselves with briefly discussing Akaike's approach and closely related criteria. Akaike's generalized likelihood function leads to the expression

$$\mathrm{AIC}(m) = n \log \widehat{\sigma}^2(m) + 2m, \tag{1.4}$$

where n is the sample size and $\widehat{\sigma}^2(m)$ is as in (1.2). An estimator for the order q is then obtained by minimizing $\mathrm{AIC}(m)$, $m \in \{0, 1, \dots, K\}$, for some predefined $0 \le q \le K$. Consistent modifications are obtained by inserting an increasing sequence C_n, and $\mathrm{AIC}(m)$ then becomes

$$\widetilde{\mathrm{AIC}}(m) = n \log \widehat{\sigma}^2(m) + 2C_n m, \qquad m \in \{0, 1, \dots, K\}. \tag{1.5}$$

Most modifications result in $C_n = \mathcal{O}(\log n)$, even though the arguments are sometimes quite different. A notable exception is the idea of Hannan and Quinn [25], who successfully employed the LIL to obtain $C_n = \mathcal{O}(\log \log n)$.

The aim of this paper is to introduce a different approach, based on the quantity $\max_{1 \le i \le d_n} |\widehat{\theta}_i - \theta_i|$, where d_n is an increasing function in n. It is shown, for instance, that, appropriately normalized, this expression converges weakly to a Gumbel-type distribution. On one hand, this allows to construct simultaneous confidence bands for the Yule–Walker estimators $\widehat{\mathbf{\Theta}}_{d_n}$, but also permits us to construct a variety of different, consistent estimators for the order q of an autoregressive process. The asymptotic distribution of such a particular estimator is also derived. As a byproduct, it is shown that known consistent criteria such as BIC, SIC and HQC are also consistent if the parameter space is increasing; that is, consistency even holds if $q \in \{0, \dots, d_n\}$, where $d_n = \mathcal{O}(n^{\delta})$. This partially gives answers to questions raised by Hannan and Quinn [25] and Shibata [41], and extends results given by An et al. [3]. In addition, the general method seems to be very useful for model fitting for subset autoregressive processes (see, e.g., [33]), which is highlighted in Remark 2.11 and Section 3. A more thorough treatment of this issue is postponed to a subsequent paper.

2. Main results.

We will frequently use the following notation. For a vector $x = (x_1, \dots, x_d)^T$, we put $\|x\|_\infty = \max_{1 \le i \le d} |x_i|$, and for a matrix $\mathbf{A} = (a_{i,j})_{\{1 \le i \le r, 1 \le j \le s\}}$, $r, s \in \mathbb{N}$ we denote with

$$\|\mathbf{A}\|_\infty = \max\{\mathbf{A}x | x \in \mathbb{R}^s, \|x\|_\infty = 1\} = \max_{1 \le i \le r} \sum_{j=1}^{s} |a_{i,j}| \tag{2.1}$$

the usual induced matrix norm. In addition, we will use the abbreviation $\| \cdot \|_p = (\mathbb{E}(| \cdot |^p))^{1/p}$, $p < \infty$. The main results involve an array of AR(q) processes;

more precisely, we consider the family of $\mathrm{AR}(d_n)$ processes $\{X_k^{(r)}\}_{k\in\mathbb{Z}}$, $1 \leq r \leq d_n$, where $d_n = \mathcal{O}(n^\delta)$ (more details are given later). Since we are always only dealing with a single member of this array, the index (r) is dropped for convenience, and we just consider an $\mathrm{AR}(d_n)$ process $\{X_k\}_{k\in\mathbb{Z}}$, keeping in mind that the parameters $\{\theta_i\}_{1\leq i\leq d_n}$ may depend on n. This implies that X_k satisfies the recurrence relation

$$X_k = \theta_1 X_{k-1} + \cdots + \theta_{d_n} X_{k-d_n} + \varepsilon_k, \qquad k \in \mathbb{Z}, \tag{2.2}$$

where $\{\varepsilon_k\}_{k\in\mathbb{Z}}$ defines the usual innovations. Note that d_n does not need to reflect the actual order q of the $\mathrm{AR}(d_n)$ process, as we do not require that $\{\theta_i\}_{1\leq i\leq d_n}$ are all different from zero. All of the results are derived under the following assumption regarding the $\mathrm{AR}(d_n)$ process $\{X_k\}_{k\in\mathbb{Z}}$.

ASSUMPTION 2.1. $\{X_k\}_{k\in\mathbb{Z}}$ admits a causal representation $X_k = \sum_{i=0}^{\infty} \alpha_i \times \varepsilon_{k-i}$, such that:

- $\sup_n \Psi(m) = \mathcal{O}(m^{-\vartheta})$, $\vartheta > 0$, where $\Psi(m) := \sum_{i=m}^{\infty} |\alpha_i|$,
- $\{\varepsilon_k\}_{k\in\mathbb{Z}}$ is a mean-zero i.i.d. sequence of random variables, such that $\|\varepsilon_k\|_p < \infty$ for some $p > 4$, $\|\varepsilon_k\|_2^2 = \sigma^2 > 0$, $k \in \mathbb{Z}$,
- $\sup_n \sum_{i=1}^{\infty} |\theta_i| < \infty$, $|\theta_n| = o((\log n)^{-1})$.

In accordance with the previously established notation, we introduce the inverse and estimated inverse matrix

$$\mathbf{\Gamma}_{d_n}^{-1} = (\gamma_{i,j}^*)_{1\leq i,j\leq d_n}, \qquad \widehat{\mathbf{\Gamma}}_{d_n}^{-1} = (\widehat{\gamma}_{i,j}^*)_{1\leq i,j\leq d_n}. \tag{2.3}$$

In addition, we will use the convention that $\theta_0 = \widehat{\theta}_0 = -1$. We can now formulate our main result.

THEOREM 2.2. *Let $\{X_k\}_{k\in\mathbb{Z}}$ be an $\mathrm{AR}(d_n)$ process satisfying Assumption 2.1. Suppose that $d_n \to \infty$ as n increases, with $d_n = \mathcal{O}(n^\delta)$ such that*

$$0 < \delta < \min\{1/2, \vartheta p/2\}, \qquad (1 - 2\vartheta)\delta < (p-4)/p. \tag{2.4}$$

If we have in addition that $\inf_h |\gamma_{h,h}^*| > 0$, *then for $z \in \mathbb{R}$*

$$P\Big(a_n^{-1}\Big(\sqrt{n} \max_{1\leq i\leq d_n} |(\widehat{\gamma}_{i,i}^* \widehat{\sigma}^2(d_n))^{-1/2}(\widehat{\theta}_i - \theta_i)| - b_n\Big) \leq z\Big) \to \exp(-e^{-z}),$$

where $a_n = (2\log d_n)^{-1/2}$ and $b_n = (2\log d_n)^{1/2} - (8\log d_n)^{-1/2}(\log\log d_n + 4\pi - 4)$.

REMARK 2.3. Condition $\inf_h |\gamma_{h,h}^*| > 0$ may be explicitly expressed in terms of $\{\theta_i\}_{1\leq i\leq d_n}$ [see (4.1)], and is quite general. In fact, it is only needed to control or exclude possible pathological cases.

REMARK 2.4. Note that if we have $|\alpha_i| = \mathcal{O}(i^{-3/2})$, then $\vartheta \geq 1/2$. Hence condition $p > 4$ implies that we may choose δ arbitrarily close to $1/2$, which essentially results in $d_n = \mathcal{o}(\sqrt{n})$.

The above remark indicates that we may obtain simple bounds for d_n, provided that we can control α_i asymptotically. If the cardinality of the set $\{1 \leq i \leq d_n | \theta_i \neq 0\}$ tends to infinity as n increases, then establishing general and simple conditions on the relation between $\{\theta_i\}_{1\leq i\leq d_n}$ and $\Psi(m)$ seems to be very difficult. One may, however, obtain the following corollary.

COROLLARY 2.5. *Suppose that* $\{\varepsilon_k\}_{k\in\mathbb{Z}}$ *is a mean-zero i.i.d. sequence of random variables, such that* $\|\varepsilon_k\|_p < \infty$ *for some* $p > 4$, $\|\varepsilon_k\|_2^2 = \sigma^2 > 0$, $k \in \mathbb{Z}$, *and that one of the following conditions holds*:

(i) $\sup_n \sum_{i=1}^{d_n} |\theta_i| < 1$,
(ii) $\theta_i = 0$, $q < i \leq d_n$ *for some fixed* $q \in \mathbb{N}$ *which does not depend on* n.

Then the conditions of Theorem 2.2 are satisfied and we can choose any $d_n = \mathcal{O}(n^\delta)$ *with* $\delta < 1/2$.

REMARK 2.6. The rate of convergence to an extreme-value type distribution as given in Theorem 2.2 can be rather slow; see, for instance, [5, 35]. Hence, in view of (1.3) (and Theorem 6.1), it may be more appropriate to use the approximation

$$P\Big(\max_{1\leq i\leq n} \big|\sqrt{n}(\widehat{\gamma}_{i,i}^{*}\widehat{\sigma}^2)^{-1/2}(\widehat{\theta}_i - \theta_i)\big| \leq x\Big) \approx P(\|\boldsymbol{\xi}_{d_n}\|_\infty \leq x)$$

in practice, where $\boldsymbol{\xi}_{d_n} = (\xi_{n,1}, \ldots, \xi_{n,d_n})^T$ is a d_n-dimensional mean-zero Gaussian random vector with the same covariance structure. Corresponding quantiles can be obtained, for instance, via a Monte Carlo technique. However, if d_n is sufficiently large, one has that

$$P(\|\boldsymbol{\xi}_{d_n}\|_\infty \leq x) \approx P(\|\boldsymbol{\eta}_{d_n}\|_\infty \leq x),$$

where $\boldsymbol{\eta}_{d_n} = (\eta_{n,1}, \ldots, \eta_{n,d_n})^T$ is a sequence of i.i.d. mean-zero Gaussian random variables with unit variance. A bound for the error can be given by using the techniques developed by Berman [9] and Deo [18]; see also the proof of Theorem 2.8.

The above results allow us to construct the simultaneous confidence bands

$$\begin{aligned} \mathcal{M}_1(d_n) = \Big\{&\boldsymbol{\Theta}_{d_n} \in \mathbb{R}^{d_n} | \\ &a_n^{-1}\Big(\sqrt{n}\max_{1\leq i\leq d_n} |(\widehat{\gamma}_{i,i}^{*})^{-1/2}(\widehat{\theta}_i - \theta_i)| - b_n\Big) \leq \sqrt{\widehat{\sigma}^2(d_n)}V_{1-\alpha}\Big\}, \end{aligned} \tag{2.5}$$

where $V_{1-\alpha}$ denotes the $1-\alpha$ quantile of the Gumbel-type distribution given above. In the literature [4, 17, 23] one often finds the confidence ellipsoids

$$\begin{aligned}\mathcal{M}_2(m) = \{&\mathbf{\Theta}_m \in \mathbb{R}^m | \\ &(\widehat{\mathbf{\Theta}}_m - \mathbf{\Theta}_m)\widehat{\mathbf{\Gamma}}_m(\widehat{\mathbf{\Theta}}_m - \mathbf{\Theta}_m)^T \le n^{-1}\widehat{\sigma}^2(m)\chi^2_{1-\alpha}(m)\},\end{aligned} \tag{2.6}$$

where $\chi^2_{1-\alpha}(m)$ denotes the $1-\alpha$ quantile of the chi-squared distribution with m degrees of freedom. Note that in general $\mathcal{M}_1(d_n) \nsubseteq \mathcal{M}_2(d_n)$ and vice versa. The confidence region $\mathcal{M}_2(d_n)$ can be viewed as a global measure, where the impact of single elements $\{|\widehat{\theta}_i - \theta_i|\}_{1\le i\le d_n}$ is negligible, which in turn leads to suboptimal confidence regions for single elements. In contrast, $\mathcal{M}_1(d_n)$ can be viewed as a local measure where single elements have a large impact, which clearly leads to significantly tighter bounds for the single elements $\{|\widehat{\theta}_i - \theta_i|\}_{1\le i\le d_n}$. This is a very important issue for so-called *subset autoregressive* models; see Remark 2.11.

Theorem 2.2 not only can be used to construct simultaneous confidence bands for the Yule–Walker estimators $\widehat{\mathbf{\Theta}}_{d_n}$, but also provides a test for the degree of an AR(q)-process. To be more precise, for an AR(q)-process $\{X_k\}_{k\in\mathbb{Z}}$ satisfying the assumptions of Theorem 2.2, we formulate the null hypothesis $\mathcal{H}_0 : q \le q_0$, and the alternative $\mathcal{H}_A : q > q_0$. Since for any fixed $k \ge 1$

$$P\Big(a_n^{-1}\Big(\sqrt{n}\max_{1\le i\le k}|(\widehat{\gamma}^*_{i,i}\widehat{\sigma}^2(d_n))^{-1/2}\widehat{\theta}_i| - b_n\Big) \le z\Big) \to 1$$

as n increases, it follows immediately from Theorem 2.2 that under $\mathcal{H}_0$ we have

$$P\Big(a_n^{-1}\Big(\sqrt{n}\max_{q_0+k\le i\le d_n}|(\widehat{\gamma}^*_{i,i}\widehat{\sigma}^2(d_n))^{-1/2}\widehat{\theta}_i| - b_n\Big) \le z\Big) \to \exp(-e^{-z})$$

for any fixed integer $k \ge 1$, since we are assuming that $\theta_i = 0$ for $i > q_0$. Conversely, it is not hard to verify (see the proof of Theorem 2.8 for details) that the quantity

$$a_n^{-1}\Big(\sqrt{n}\max_{q_0+1\le i\le d_n}|(\widehat{\gamma}^*_{i,i}\widehat{\sigma}^2(d_n))^{-1/2}\widehat{\theta}_i| - b_n\Big)$$

explodes under the alternative $\mathcal{H}_A : q > q_0$. This can be used to establish a lower bound for the order q or to test if the order was chosen sufficiently large. This is particularly useful if q is large compared to the sample size and the magnitude of $\mathbf{\Theta}_q$, in which case the AIC and related criteria sometimes heavily fail to get near the true order. More details on this subject and examples are given in Section 3. Generally speaking, such situations are often encountered in *subset autoregressive* models; see Remark 2.11.

The above conclusions lead to the following family of estimators $\widehat{q}^{(1)}_{z_n}$ for q. Let z_n be a monotone sequence that tends to infinity as n increases. Then we define the estimator

$$\widehat{q}^{(1)}_{z_n} = \min\Big\{q \in \mathbb{N} | a_n^{-1}\Big(\sqrt{n}\max_{q+1\le i\le d_n}|(\widehat{\gamma}^*_{i,i}\widehat{\sigma}^2(d_n))^{-1/2}\widehat{\theta}_i| - b_n\Big) \le z_n\Big\}. \tag{2.7}$$

Using the above ideas, it is not hard to show that the estimators $\widehat{q}_{z_n}^{(1)}$ are consistent if z_n does not grow too fast. In fact, under some more conditions imposed on the sequence z_n, we can even derive the asymptotic distribution of the estimators.

ASSUMPTION 2.7. In addition to Assumption 2.1, suppose that:

- $\sum_{i=1}^{\infty} |\theta_i| < \infty$, $|\theta_n| = \mathcal{O}((\log n)^{-2-\eta})$, $\eta > 0$,
- $\mathbb{E}(\exp(\lambda |\varepsilon_k|)) < \infty$, for some $\lambda > 0$ and all $k \in \mathbb{Z}$,
- $|\alpha_i| = \mathcal{O}(i^{-\beta})$, $\beta > 3/2$.

THEOREM 2.8. *Let $\{X_k\}_{k\in\mathbb{Z}}$ be an* AR(q)*-process such that Assumption 2.7 is valid. Assume in addition that* $\inf_h |\gamma_{h,h}^*| > 0$ *and* $z_n = \mathcal{O}(\log n)$. *Then if* $z_n \to \infty$, *the estimator* $\widehat{q}_{z_n}^{(1)}$ *in* (2.7) *is consistent. Moreover, the following expansion is valid*:

$$P\big(\widehat{q}_{z_n}^{(1)} = k + q\big) = \frac{e^{-z_n}}{d_n} + o\left(\frac{e^{-z_n}}{d_n} + d_n^{-z_n^2+1}\right)$$

for $k \in \mathbb{N}$, $k = \mathcal{O}(n^{\delta})$, $\delta < 1/7$.

REMARK 2.9. The stronger conditions of Assumption 2.7 are necessary to control the rate of convergence in Theorem 2.2, which in turn allows for the explicit expansion given above. This, however, also leads to the more restrictive bound $q + k = \mathcal{O}(d_n) = \mathcal{O}(n^{\delta})$, $\delta < 1/7$; see also Remark 6.2. If we are only interested in establishing consistency, then we may drop these more restrictive assumptions; see in particular Theorem 2.12 below.

REMARK 2.10. Theorem 2.8 yields that in some sense the estimators $\widehat{q}_{z_n}^{(1)}$ possess a discrete uniform asymptotic distribution, which leads to the surprising conclusion

$$P\big(\widehat{q}_{z_n}^{(1)} = 1 + q\big) \approx P\big(\widehat{q}_{z_n}^{(1)} = 1000 + q\big).$$

This fact can be explained by the maximum function in the definition of $\widehat{q}_{z_n}^{(1)}$, more precisely, due to the weak dependence of the Yule–Walker estimators $\widehat{\mathbf{\Theta}}_{d_n}$. The maximum function essentially does not care at which index i the boundary z_n is exceeded, and this results in the uniform distribution. It turns out (see Section 3) that a modified version of the estimator $\widehat{q}_{z_n}^{(1)}$ is a very efficient preliminary estimator that establishes a decent lower bound.

An asymptotic uniform-type distribution clearly is not a desirable property for an estimator. However, similarly to Akaike's method, we can introduce a penalty function and construct different yet also consistent estimators for the order q.

To this end, for $x \in \mathbb{R}$ put $(x)^+ = \max(0, x)$ and let $\Upsilon_{n,i} = a_n^{-1}(\sqrt{n}|(\widehat{\gamma}_{i,i}^* \times \widehat{\sigma}^2(d_n))^{-1/2}\widehat{\theta}_i| - b_n)$. Then we introduce a new estimator $\widehat{q}_{z_n}^{(2)}$ as

$$\widehat{q}_{z_n}^{(2)} = \arg\min_{q\in\mathbb{N}}\Big\{\max_{q+1\le i\le d_n}\{(\Upsilon_{n,i} - z_n)^+\} + \log(1+q)\Big\}.$$

More generally, let $\mathcal{F} = (f_d)_{d\in\mathbb{N}}$ be a collection of continuous functions such that:

- f_d is a map from $\mathbb{R}^{d+2}$ to $\mathbb{R}$,
- $f_d(0, \dots, 0, q, d) < f_d(0, \dots, 0, q+1, d)$ for all $d, q \in \mathbb{N}$,
- if $a_n, d_n \to \infty$ as n increases, then $f_{d_n}(\dots, a_n, \dots, q, d_n) \to \infty$ as n increases, regardless of the values of the other coordinates.

Define

$$\widehat{q}_{z_n}^{(f)} = \arg\min_{q\in\mathbb{N}} f_{d_n}\big(0, \dots, 0, (\Upsilon_{n,q+1} - z_n)^+, \dots, (\Upsilon_{n,d_n} - z_n)^+, q, d_n\big). \tag{2.8}$$

Then arguing as in the proof of Theorem 2.8 it can be shown that this constitutes a consistent estimator for the true value q. For example, the following estimator

$$\widehat{q}_{z_n}^{(3)} = \arg\min_{q\in\mathbb{N}}\Big\{\sum_{q+1\le i\le d_n}(\Upsilon_{n,i} - z_n)^+ + q\Big\}$$

satisfies the conditions above and is consistent.

REMARK 2.11. Note that instead of defining a specific order q, one can also consider a special lag configuration, for example, $\mathbf{\Theta}_q = (\theta_1, \theta_2, 0, \dots, 0, \theta_{10}, \theta_{11}, \dots, \theta_q)^T$. Such configurations are commonly referred to as *subset autoregressive models*; see, for instance, [16, 32, 33, 42, 46] and the references therein. The AIC(m) and especially related consistent criteria have problems dealing with such *subset autoregressive models*, which can be seen as follows. By Hannan [23], Chapter VI, we have for $m \in \mathbb{N}$

$$\begin{aligned}\widetilde{\mathrm{AIC}}(m)n^{-1} &= \log(\widehat{\sigma}^2(m)) + 2n^{-1}C_n m\\ &= \log\widehat{\phi}_{n,0} + \sum_{i=1}^{m}\log\big(1 - \widehat{\theta}_i^2(m)\big) + 2n^{-1}C_n m.\end{aligned} \tag{2.9}$$

This shows that in case of *subset autoregressive models*, the penalty function $2n^{-1}C_n m$ is too severe and should be replaced, at least in theory, by $2n^{-1}C_n \times \sum_{i=1}^{m}\mathbf{1}_{\{\theta_i\neq 0\}}$, since this is impossible in practice. Of course the same problem arises if some of the $\{\theta_i\}_{1\le i\le q}$ are close to zero. A maximum based estimator like $\widehat{q}_{z_n}^{(1)}$ gets less effected, which is empirically confirmed in Section 3.

An often encountered theoretical assumption for estimators related to AIC(m) is that the parameter space for q is finite; that is, it is usually assumed in advance that $q \in \{0, \dots, K\}$, where K is "chosen sufficiently large," but finite. In [25], K is allowed to increase with the sample size with unknown rate, which was specified later by An et al. [3]. Note, however, that for the estimators defined above we allow $K = K_n = d_n$. Before extending this result, we give precise definitions of BIC, HQC, MIC (=miscellaneous information criterion) and SIC, as the literature does not seem to be very clear on this subject, in particular in the case of the BIC and SIC. In the sequel, the following definitions are used:

$$\begin{aligned} \mathrm{BIC}(m) &= \mathrm{SIC}(m) = \log\widehat{\sigma}^2(m) + mn^{-1}\log n, \\ \mathrm{MIC}(m) &= \log\widehat{\sigma}^2(m) + m/2n^{-1}\log n, \\ \mathrm{HQC}(m) &= \log\widehat{\sigma}^2(m) + n^{-1}2cm\log\log n, \qquad c > 1. \end{aligned} \tag{2.10}$$

This means that we use the same definitions for BIC and SIC (asymptotically), which is the case mostly encountered in the literature. The MIC differs from the BIC by the choice of the constant $1/2$ that naturally leads to a less parsimonious criterion, which performs quite well in the examples given in Section 3. Using some of the results of Section 4 and 6, one may prove the following.

THEOREM 2.12. *Assume that the conditions of Theorem 2.2 hold, and additionally assume that* $\inf_h |1 - \theta_h^2| > 0$. *Let C_n be a positive sequence such that*:

- $\lim_n C_n(2\log\log n)^{-1} > 1$, $C_n = o(n)$,
- $\log d_n \leq o(C_n)$.

Then the estimators for the order q defined as

$$\widehat{q}_n^* = \underset{0\leq m\leq d_n}{\arg\min}\bigl(\log\widehat{\sigma}^2(m) + n^{-1}C_n m\bigr)$$

are consistent.

REMARK 2.13. Note that condition $\inf_h|1 - \theta_h^2| > 0$ essentially is already provided by the causality condition in Assumption 2.1.

Theorem 2.12 thus implies the bounds $d_n \in \mathcal{O}\{o(n^{1/2}), o(n^{1/2}), o(\log n)\}$ for BIC, MIC and HQC, and thus significantly improves the bounds provided by An et al. [3] [BIC: $\mathcal{O}(\log n)$, HQC: $\mathcal{O}(\log\log n)$]. On the other hand, the setting in An et al. [3] is more general, and it is also shown that the estimators are strongly consistent.

3. Simulation and numerical results. In this section we will perform a small simulation study to compare some of the previously mentioned estimators. We will look at the performance in case of AR(6), AR(12) and AR(24) processes. The sample size n satisfies $n \in \{125, 250, 500, 1000\}$; as for the dimension d_n, we chose the functions $d_n \in \{2\log n, 4\log n, 6\log n\}$, and rounded up the values. This implies that the parameter space $q \in \{0, \dots, K\}$ satisfies $K \in \{10, 12, 13, 14\}$, $K \in \{20, 23, 25, 29\}$, $K \in \{29, 34, 38, 42\}$. For reference, note that $\{\lceil\sqrt{125}\rceil, \lceil\sqrt{250}\rceil, \lceil\sqrt{500}\rceil, \lceil\sqrt{1000}\rceil\} = \{12, 16, 23, 32\}$. To introduce the estimators $\widehat{q}_{z_n}^{(4)}(d_n), \widehat{q}_{z_n}^{(5)}(d_n)$, we require some additional notation. For $1 \le k \le d_n$, define $\{\widehat{\gamma}_{i,i}^*(k)\}_{1\le i\le k}$ and $\{\widehat{\theta}_i(k)\}_{1\le i\le k}$ via the usual relation

$$\widehat{\Theta}_k = \widehat{\mathbf{\Gamma}}_k^{-1}\widehat{\Phi}_k. \tag{3.1}$$

The estimators are now defined as

$$\widehat{q}_{z_n}^{(4)}(k) = \min\Bigl\{q \in \mathbb{N} \big| a_n^{-1}\Bigl(\sqrt{n} \max_{q+1\le i\le k} |(\widehat{\gamma}_{i,i}^*(k)\widehat{\sigma}^2(k))^{-1/2}\widehat{\theta}_i(k)| - b_n\Bigr) \le z_n\Bigr\},$$

$$\widehat{q}_{z_n}^{(5)}(d_n) = \max_{1\le k\le d_n} \widehat{q}_{z_n}^{(4)}(k).$$

Note that the definition of a_n, b_n remains unchanged. This modification significantly improves the performance in practice, which is due to the following reason: if one just considers the estimator $\widehat{q}_{z_n}^{(4)}(d_n)$ and hence *only the equation* $\widehat{\Theta}_{d_n} = \widehat{\mathbf{\Gamma}}_{d_n}^{-1}\widehat{\Phi}_{d_n}$, the bias may be quite large since the estimate $\widehat{\mathbf{\Gamma}}_{d_n}^{-1}$ is rather poor for larger d_n. Note that this is also true when computing the AIC or related criteria, which is a well-established fact in the literature (cf. [2, 17, 23, 25]). Hence one may expect that the "maximum" version $\widehat{q}_{z_n}^{(5)}(d_n)$ outperforms its counterpart $\widehat{q}_{z_n}^{(4)}(d_n)$, which is indeed the case in the examples given below. The values for z_n were chosen as $z_n \in \{x_n, y_n\}$, where x_n satisfies $a_n x_n + b_n = 2.71$ for $n \in \{125, 250\}$, $a_n x_n + b_n = 2.91$ for $n \in \{500, 1000\}$. Similarly, we have $a_n y_n + b_n = 3$ for $n \in \{125, 250\}$, $a_n y_n + b_n = 3.2$ for $n \in \{500, 1000\}$. This means that the estimators get less parsimonious when d_n increases. Of course an adaption to maintain the same confidence level is possible, but the general picture remains the same.

For the criteria AIC, BIC, HQC and MIC we use the definitions given in (1.4) and (2.10); in case of HQC we choose $c = 1$, since, as pointed out by Hannan and Quinn [25], "it would seem pedantic to choose values as $c = 1.01$." The following modifications are also considered:

$$\begin{aligned} \mathrm{AIC}(m)^* &= \max\{\mathrm{AIC}(m), \widehat{q}_{y_n}^{(5)}(d_n)\},\\ \mathrm{BIC}(m)^* &= \max\{\mathrm{BIC}(m), \widehat{q}_{y_n}^{(5)}(d_n)\},\\ \mathrm{HQC}(m)^* &= \max\{\mathrm{HQC}(m), \widehat{q}_{y_n}^{(5)}(d_n)\},\\ \mathrm{MIC}(m)^* &= \max\{\mathrm{MIC}(m), \widehat{q}_{y_n}^{(5)}(d_n)\}. \end{aligned} \tag{3.2}$$

All simulations were carried out using the program R;[1] in order to get a sample of size n, a sample path of size $1000 + n$ was produced and the first 1000 observations were discarded.

Generally speaking, unreported simulations show that in many cases the modified criteria $\mathrm{AIC}(m)^*, \mathrm{BIC}(m)^*, \dots$ perform nearly identically as the nonmodified

[1] http://portal.tugraz.at/portal/page/portal/TU_Graz/Einrichtungen/Institute/Homepages/i5060/research/R_Code.

ones $\mathrm{AIC}(m), \mathrm{BIC}(m), \ldots$. This is in particular the case when dealing with full parameter sets, that is, $\theta_i \neq 0$, $1 \leq i \leq q$, and θ_q is sufficiently large. If this is the case, the performance of the estimators $\widehat{q}^{(5)}_{x_n}(d_n), \widehat{q}^{(5)}_{y_n}(d_n)$ is somewhere between the $\mathrm{BIC}(m)$ and $\mathrm{HQC}(m)$. On the other hand, if the model is not full and/or the order q is sufficiently large, then the differences can be quite striking. The aim of the following examples is to illustrate this behavior.

3.1. AR(6). First note that the definitions of x_n, y_n result in

$$P(\max|\boldsymbol{\xi}| \leq 2.71) \geq 0.92, \qquad P(\max|\boldsymbol{\xi}| \leq 3) \geq 0.97, \qquad d_n \in \{10, 12\},$$

$$P(\max|\boldsymbol{\xi}| \leq 2.91) \geq 0.95, \qquad P(\max|\boldsymbol{\xi}| \leq 3.2) \geq 0.98, \qquad d_n \in \{13, 14\},$$

where $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_{d_n})^T$ is a d_n-dimensional mean-zero Gaussian random vector where the covariance matrix is the identity.

The results shown in Tables 1 and 2 hint at what is to be expected in case of full models, namely that the modifications $\mathrm{AIC}(m)^*, \mathrm{BIC}(m)^*, \ldots$ perform nearly as well as the normal versions $\mathrm{AIC}(m), \mathrm{BIC}(m), \ldots$. The estimators $\widehat{q}^{(5)}_{x_n}(d_n)$, $\widehat{q}^{(5)}_{y_n}(d_n)$ perform also quite well.

Contrary to the previous results, Tables 3 and 4 show the difference of the modified estimators [and $\widehat{q}^{(5)}_{x_n}(d_n), \widehat{q}^{(5)}_{y_n}(d_n)$], if the model is very sparse. Except for the case $n = 1000$, the modifications are notably better.

3.2. AR(12). The definitions of x_n, y_n result in

$$P(\max|\boldsymbol{\xi}| \leq 2.71) \geq 0.85, \qquad P(\max|\boldsymbol{\xi}| \leq 3) \geq 0.94, \qquad d_n \in \{20, 23\},$$

$$P(\max|\boldsymbol{\xi}| \leq 2.91) \geq 0.9, \qquad P(\max|\boldsymbol{\xi}| \leq 3.2) \geq 0.96, \qquad d_n \in \{25, 29\},$$

TABLE 1
Simulation of an AR(6) *process with coefficients* $\Theta_6 = (0.1, -0.3, 0.05, 0.2, -0.1, 0.2)^T$, $\varepsilon \sim \mathcal{N}(0, 1)$, 1000 *repetitions*, $d_n \in \{10, 12\}$

n	$\widehat{q}$	**AIC**	**AIC***	**BIC**	**BIC***	**HQC**	**HQC***	**MIC**	**MIC***	$\widehat{q}^{(5)}_{y_n}$	$\widehat{q}^{(5)}_{x_n}$
125	<5	428	427	943	808	746	704	550	545	816	701
	5	65	65	10	30	32	40	58	58	28	41
	6	**344**	**341**	**45**	**143**	**191**	**214**	**295**	**294**	**137**	**196**
	7	66	65	1	5	23	24	54	53	5	14
	<7	97	102	1	14	8	18	43	50	14	48
250	<5	93	89	693	432	328	282	202	188	440	299
	5	24	23	14	32	32	32	33	31	42	38
	6	**646**	**632**	**287**	**481**	**586**	**595**	**649**	**634**	**467**	**543**
	7	96	95	5	8	37	35	74	73	4	9
	>7	141	161	1	47	17	56	42	74	47	111

TABLE 2
Simulation of an AR(6) *process with coefficients* $\Theta_6 = (0.1, -0.3, 0.05, 0.2, -0.1, 0.2)^T$, $\varepsilon \sim \mathcal{N}(0, 1)$, 1000 *repetitions*, $d_n \in \{13, 14\}$

n	$\widehat{q}$	**AIC**	**AIC***	**BIC**	**BIC***	**HQC**	**HQC***	**MIC**	**MIC***	$\widehat{q}^{(5)}_{y_n}$	$\widehat{q}^{(5)}_{x_n}$
500	<5	1	1	177	75	29	25	15	15	86	52
	5	3	3	9	11	6	6	3	3	17	14
	6	**730**	**713**	**805**	**874**	**913**	**889**	**892**	**867**	**865**	**849**
	7	108	108	8	8	42	42	57	57	0	2
	<7	158	175	1	32	10	38	33	58	32	83
1000	<5	0	0	3	0	0	0	0	0	0	0
	5	0	0	0	0	0	0	0	0	0	0
	6	**724**	**709**	**990**	**951**	**952**	**917**	**934**	**901**	**955**	**885**
	7	103	101	7	9	36	34	47	44	5	7
	>7	173	190	0	40	12	49	19	55	40	108

where $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_{d_n})^T$ is a d_n-dimensional mean-zero Gaussian random vector where the covariance matrix is the identity.

The results are depicted in Tables 5, 6, 7 and 8, and are quite similar to the case of the AR(6) processes. If the model is rather full, AIC$(m)^*$, BIC$(m)^*$, ... perform nearly as well as the normal versions AIC(m), BIC(m), ..., whereas in case of the sparse model, a significant difference can be observed.

3.3. AR(24). In this case, the definitions of x_n, y_n result in

$$P(\max|\boldsymbol{\xi}| \leq 2.71) \geq 0.795, \qquad P(\max|\boldsymbol{\xi}| \leq 3) \geq 0.912, \qquad d_n \in \{29, 34\},$$

TABLE 3
Simulation of an AR(6) *process with coefficients* $\Theta_6 = (0.1, 0, 0.05, 0, 0, 0.2)^T$, $\varepsilon \sim \mathcal{N}(0, 1)$, 1000 *repetitions*, $d_n \in \{10, 12\}$

n	$\widehat{q}$	**AIC**	**AIC***	**BIC**	**BIC***	**HQC**	**HQC***	**MIC**	**MIC***	$\widehat{q}^{(5)}_{y_n}$	$\widehat{q}^{(5)}_{x_n}$
125	<5	719	699	998	854	944	842	839	787	854	747
	5	11	11	0	0	2	2	7	7	0	11
	6	**168**	**181**	**2**	**124**	**43**	**126**	**107**	**145**	**124**	**184**
	7	44	44	0	4	8	11	23	24	4	8
	<7	58	65	0	18	3	19	24	37	18	50
250	<5	290	276	960	437	723	424	550	396	438	321
	5	6	6	0	3	2	3	5	5	3	5
	6	**491**	**488**	**39**	**513**	**245**	**503**	**376**	**494**	**513**	**573**
	7	91	90	1	2	21	21	40	40	1	7
	>7	122	140	0	45	9	49	29	65	45	94

TABLE 4
Simulation of an AR(6) *process with coefficients* $\Theta_6 = (0.1, 0, 0.05, 0, 0, 0.2)^T$, $\varepsilon \sim \mathcal{N}(0, 1)$, 1000 *repetitions*, $d_n \in \{13, 14\}$

n	$\widehat{q}$	AIC	AIC*	BIC	BIC*	HQC	HQC*	MIC	MIC*	$\widehat{q}^{(5)}_{y_n}$	$\widehat{q}^{(5)}_{x_n}$
500	<5	21	21	761	102	267	98	164	85	102	56
	5	0	0	1	0	0	0	0	0	0	1
	6	**663**	**655**	**234**	**871**	**675**	**822**	**736**	**796**	**874**	**863**
	7	125	124	4	3	50	49	69	68	0	10
	<7	191	200	0	24	8	31	31	51	24	70
1000	<5	0	0	168	1	3	1	1	1	1	0
	5	0	0	0	0	0	0	0	0	0	0
	6	**702**	**683**	**822**	**949**	**940**	**905**	**919**	**887**	**955**	**898**
	7	121	119	9	9	43	42	52	52	3	9
	>7	177	198	1	41	14	52	28	60	41	93

$$P(\max|\boldsymbol{\xi}| \leq 2.91) \geq 0.86, \qquad P(\max|\boldsymbol{\xi}| \leq 3.2) \geq 0.94, \qquad d_n \in \{38, 42\},$$

where $\boldsymbol{\xi} = (\xi_1, \dots, \xi_{d_n})^T$ is a d_n-dimensional mean-zero Gaussian random vector where the covariance matrix is the identity. The behavior shown in Tables 9, 10, 11 and 12 is as in the previous two cases. The difference in the sparse model is perhaps the most striking one.

4. Proofs and ramification. In this section, we will prove Theorems 2.2, 2.8, 2.12, and also explicitly mention some auxiliary results which have interest in themselves. For $d_n \leq m$ let $\boldsymbol{\Gamma}_m^{-1} = (\gamma^*_{i,j})_{1 \leq i,j \leq m}$ be the inverse of the covariance

TABLE 5
Simulation of an AR(12) *process with nonzero coefficients* $\theta_1 = 0.1$, $\theta_3 = -0.4$, $\theta_5 = 0.5$, $\theta_7 = -0.1$, $\theta_8 = 0.05$, $\theta_{10} = -0.3$, $\theta_{12} = 0.2$, $\varepsilon \sim \mathcal{N}(0, 1)$, 1000 *repetitions*, $d_n \in \{20, 23\}$

n	$\widehat{q}$	AIC	AIC*	BIC	BIC*	HQC	HQC*	MIC	MIC*	$\widehat{q}^{(5)}_{y_n}$	$\widehat{q}^{(5)}_{x_n}$
125	<11	705	701	995	966	931	917	812	807	969	929
	11	79	79	2	3	22	22	54	54	1	2
	12	**141**	**141**	**3**	**23**	**40**	**47**	**97**	**98**	**22**	**47**
	13	48	48	0	4	6	9	30	30	4	11
	>13	27	31	0	4	1	5	7	11	4	11
250	<11	257	257	854	730	573	560	423	421	748	620
	11	39	39	9	10	31	31	39	39	3	11
	12	**495**	**493**	**135**	**247**	**349**	**356**	**442**	**441**	**237**	**313**
	13	115	115	2	4	40	40	65	65	3	13
	>13	94	96	0	9	7	13	31	34	9	43

TABLE 6

Simulation of an AR(12) *process with nonzero coefficients* $\theta_1 = 0.1$, $\theta_3 = -0.4$, $\theta_5 = 0.5$, $\theta_7 = -0.1$, $\theta_8 = 0.05$, $\theta_{10} = -0.3$, $\theta_{12} = 0.2$, $\varepsilon \sim \mathcal{N}(0, 1)$, 1000 *repetitions*, $d_n \in \{25, 28\}$

n	$\widehat{q}$	**AIC**	**AIC***	**BIC**	**BIC***	**HQC**	**HQC***	**MIC**	**MIC***	$\widehat{q}^{(5)}_{y_n}$	$\widehat{q}^{(5)}_{x_n}$
500	<11	19	19	367	256	110	106	75	73	269	183
	11	4	4	4	4	6	6	6	6	2	2
	12	**684**	**680**	**618**	**705**	**808**	**793**	**808**	**797**	**702**	**758**
	13	129	128	10	12	63	62	78	76	4	8
	>13	164	169	1	23	13	33	33	48	23	49
1000	<11	0	0	11	2	0	0	0	0	2	1
	11	0	0	0	0	0	0	0	0	0	0
	12	**679**	**676**	**970**	**947**	**925**	**900**	**896**	**873**	**958**	**914**
	13	151	150	17	17	61	60	79	78	6	13
	>13	170	174	2	34	14	40	25	49	34	72

matrix $\mathbf{\Gamma}_m = (\gamma_{i,j})_{1 \le i,j \le m}$ associated to the AR(d_n) process $\{X_k\}_{k \in \mathbb{Z}}$. Due to Galbraith and Galbraith [21], it holds that

$$\sigma^2 \gamma^*_{i,j} = \sum_{r=0}^{\alpha} \theta_r \theta_{r+j-i} - \sum_{r=\beta}^{d_n+i-j} \theta_r \theta_{r+j-i}, \qquad 1 \le i \le j \le m, \tag{4.1}$$

where

$$\alpha = \min\{i-1, d_n + i - j, m - j\}, \qquad \beta = \max\{i-1, m-j\},$$

and either of the sums is taken to be zero if its upper limit is less than its lower limit. The second sum is zero unless $m - d_n + 1 \le i \le j \le d_n$ while both sums

TABLE 7

Simulation of an AR(12) *process with nonzero coefficients* $\theta_1 = 0.1$, $\theta_3 = -0.4$, $\theta_{12} = 0.2$, $\varepsilon \sim \mathcal{N}(0, 1)$, 1000 *repetitions*, $d_n \in \{20, 23\}$

n	$\widehat{q}$	**AIC**	**AIC***	**BIC**	**BIC***	**HQC**	**HQC***	**MIC**	**MIC***	$\widehat{q}^{(5)}_{y_n}$	$\widehat{q}^{(5)}_{x_n}$
125	<10	884	853	1000	920	995	920	963	910	920	861
	11	3	3	0	0	0	0	1	1	0	3
	12	**68**	**94**	**0**	**71**	**5**	**71**	**25**	**70**	**71**	**114**
	13	11	13	0	3	0	3	4	7	3	5
	>13	34	37	0	6	0	6	7	12	6	17
250	<10	509	421	999	555	934	552	792	530	555	424
	11	3	3	0	3	0	2	2	3	3	4
	12	**340**	**419**	**1**	**421**	**59**	**419**	**170**	**416**	**421**	**514**
	13	67	68	0	2	4	6	18	19	2	5
	>13	81	89	0	19	3	21	18	32	19	53

TABLE 8
Simulation of an AR(12) *process with nonzero coefficients* $\theta_1 = 0.1$, $\theta_3 = -0.4$, $\theta_{12} = 0.2$, $\varepsilon \sim \mathcal{N}(0, 1)$, 1000 *repetitions*, $d_n \in \{25, 28\}$

n	$\widehat{q}$	**AIC**	**AIC***	**BIC**	**BIC***	**HQC**	**HQC***	**MIC**	**MIC***	$\widehat{\boldsymbol{q}}_{\boldsymbol{y_n}}^{\mathbf{(5)}}$	$\widehat{\boldsymbol{q}}_{\boldsymbol{x_n}}^{\mathbf{(5)}}$
500	<11	77	58	983	125	613	125	402	115	125	78
	11	0	0	0	2	0	2	0	1	2	1
	12	**663**	**678**	**17**	**858**	**360**	**834**	**532**	**808**	**858**	**870**
	13	104	103	0	3	15	16	39	40	3	4
	>13	156	161	0	12	12	23	27	36	12	47
1000	<11	0	0	689	2	67	2	35	2	2	2
	11	0	0	0	0	0	0	0	0	0	0
	12	**706**	**701**	**307**	**971**	**880**	**926**	**893**	**907**	**972**	**936**
	13	124	123	2	2	39	38	54	53	1	3
	>13	170	176	2	25	14	34	18	38	25	59

are zero if $j - i > d_n$. Note that this implies $\sigma^2(m)\gamma^*_{m,m} = 1$ for $m > d_n$, and in particular that

$$\sup_{|h|\geq n} \sup_i |\gamma^*_{i,i+h}| = \mathcal{O}((\log n)^{-1}), \tag{4.2}$$

if Assumption 2.1 is valid. Throughout this section and particularly in the proofs of the presented results, we use the notation $\widehat{\sigma}^2 = \widehat{\sigma}^2(d_n)$. Note that we can rewrite the equation defining the AR(d_n) process as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\Phi}_{d_n} + \mathbf{Z}, \tag{4.3}$$

TABLE 9
Simulation of an AR(24) *process with nonzero coefficients* $\theta_1 = 0.6$, $\theta_2 = -0.1$, $\theta_4 = 0.05$, $\theta_7 = 0.15$, $\theta_8 = -0.27$, $\theta_{10} = 0.1$, $\theta_{12} = -0.2$, $\theta_{15} = -0.25$, $\theta_{18} = 0.05$, $\theta_{20} = 0.1$, $\theta_{21} = -0.3$, $\theta_{24} = 0.17$, $\varepsilon \sim \mathcal{N}(0, 1)$, 1000 *repetitions*, $d_n \in \{29, 34\}$

n	$\widehat{q}$	**AIC**	**AIC***	**BIC**	**BIC***	**HQC**	**HQC***	**MIC**	**MIC***	$\widehat{\boldsymbol{q}}_{\boldsymbol{y_n}}^{\mathbf{(5)}}$	$\widehat{\boldsymbol{q}}_{\boldsymbol{x_n}}^{\mathbf{(5)}}$
125	<23	972	970	1000	996	1000	996	992	990	996	989
	23	12	12	0	1	0	1	5	5	1	2
	24	**3**	**3**	**0**	**1**	**0**	**1**	**1**	**1**	**1**	**6**
	25	10	10	0	0	0	0	2	2	0	1
	>25	3	5	0	2	0	2	0	2	2	2
250	<23	518	516	995	923	872	840	727	717	924	845
	23	120	120	2	13	48	50	77	78	12	25
	24	**185**	**186**	**3**	**57**	**67**	**90**	**135**	**138**	**57**	**98**
	25	89	89	0	1	7	8	38	38	1	10
	>25	88	89	0	6	6	12	23	29	6	22

TABLE 10

Simulation of an AR(24) *process with nonzero coefficients* $\theta_1 = 0.6$, $\theta_2 = -0.1$, $\theta_4 = 0.05$, $\theta_7 = 0.15$, $\theta_8 = -0.27$, $\theta_{10} = 0.1$, $\theta_{12} = -0.2$, $\theta_{15} = -0.25$, $\theta_{18} = 0.05$, $\theta_{20} = 0.1$, $\theta_{21} = -0.3$, $\theta_{24} = 0.17$, $\varepsilon \sim \mathcal{N}(0, 1)$, 1000 *repetitions*, $d_n \in \{38, 42\}$

n	**$\widehat{q}$**	**AIC**	**AIC***	**BIC**	**BIC***	**HQC**	**HQC***	**MIC**	**MIC***	**$\widehat{q}^{(5)}_{y_n}$**	**$\widehat{q}^{(5)}_{x_n}$**
500	<23	63	62	716	545	302	288	210	205	589	430
	23	38	38	55	60	87	87	85	85	58	71
	24	**513**	**512**	**208**	**357**	**490**	**500**	**525**	**526**	**326**	**437**
	25	192	192	18	28	93	93	129	129	19	27
	>25	194	196	3	10	28	32	51	55	8	35
1000	<23	0	0	81	30	6	5	3	3	42	18
	23	0	0	34	31	8	7	6	6	48	35
	24	**562**	**552**	**835**	**857**	**796**	**775**	**761**	**741**	**868**	**842**
	25	197	195	48	45	140	137	160	156	7	24
	>25	241	253	2	37	50	76	70	94	35	81

where $\mathbf{Y} = (X_1, \dots, X_n)^T$, $\mathbf{Z} = (\varepsilon_1, \dots, \varepsilon_n)^T$, and the $n \times d_n$ design matrix $\mathbf{X}$ is given as

$$\mathbf{X} = \begin{pmatrix} X_0 & X_{-1} & \cdots & X_{1-d_n} \\ X_1 & X_0 & \cdots & X_{2-d_n} \\ \cdots & \cdots & & \\ X_{n-1} & X_{n-2} & \cdots & X_{n-d_n} \end{pmatrix}.$$

TABLE 11

Simulation of an AR(24) *process with nonzero coefficients* $\theta_1 = 0.6$, $\theta_2 = -0.1$, $\theta_4 = 0.05$, $\theta_{10} = 0.1$, $\theta_{12} = -0.2$, $\theta_{24} = 0.17$, $\varepsilon \sim \mathcal{N}(0, 1)$, 1000 *repetitions*, $d_n \in \{29, 34\}$

n	**$\widehat{q}$**	**AIC**	**AIC***	**BIC**	**BIC***	**HQC**	**HQC***	**MIC**	**MIC***	**$\widehat{q}^{(5)}_{y_n}$**	**$\widehat{q}^{(5)}_{x_n}$**
125	<23	1000	991	1000	991	1000	991	1000	991	991	969
	23	0	2	0	2	0	2	0	2	2	6
	24	**0**	**6**	**0**	**6**	**0**	**6**	**0**	**6**	**6**	**20**
	25	0	1	0	1	0	1	0	1	1	1
	>25	0	0	0	0	0	0	0	0	0	4
250	<23	857	768	1000	817	998	817	986	815	817	702
	23	1	15	0	27	0	26	0	25	27	39
	24	**99**	**166**	**0**	**142**	**2**	**143**	**13**	**145**	**142**	**225**
	25	20	22	0	3	0	3	0	3	3	5
	>25	23	29	0	11	0	11	1	12	11	29

TABLE 12

Simulation of an AR(24) *process with nonzero coefficients* $\theta_1 = 0.6$, $\theta_2 = -0.1$, $\theta_4 = 0.05$, $\theta_{10} = 0.1$, $\theta_{12} = -0.2$, $\theta_{24} = 0.17$, $\varepsilon \sim \mathcal{N}(0, 1)$, 1000 *repetitions*, $d_n \in \{38, 42\}$

n	**$\widehat{q}$**	**AIC**	**AIC***	**BIC**	**BIC***	**HQC**	**HQC***	**MIC**	**MIC***	**$\widehat{q}^{(5)}_{y_n}$**	**$\widehat{q}^{(5)}_{x_n}$**
500	<23	351	270	1000	383	952	380	854	379	383	256
	23	2	8	0	51	0	48	0	41	51	61
	24	**451**	**522**	**0**	**547**	**45**	**550**	**130**	**545**	**547**	**637**
	25	74	73	0	0	3	3	13	13	0	2
	>25	122	127	0	19	0	19	3	22	19	44
1000	<23	10	6	986	15	440	15	280	15	15	3
	23	0	0	0	14	0	13	0	11	14	12
	24	**718**	**715**	**14**	**941**	**522**	**908**	**659**	**887**	**941**	**905**
	25	121	118	0	3	32	31	46	45	3	8
	>25	151	161	0	27	6	33	15	42	27	72

We have

$$\mathbf{\Gamma}^{-1}\mathbf{X}^T\mathbf{Z} = \mathbf{\Gamma}^{-1}\sum_{k=1}^{n}\mathbf{V}_k = \sum_{k=1}^{n}\mathbf{U}_k,$$

where $\mathbf{V}_k = (V_k^{(1)}, \dots, V_k^{(d_n)})^T$, $\mathbf{U}_k = (U_k^{(1)}, \dots, U_k^{(d_n)})^T$. The following results are key ingredients.

LEMMA 4.1. *Let* $\{X_k\}_{k\in\mathbb{Z}}$ *be an* AR(d_n) *process*, *such that Assumption* 2.1 *is valid. Then*

$$P\bigl(\|\mathbf{\Gamma}_{d_n}^{-1} - \widehat{\mathbf{\Gamma}}_{d_n}^{-1}\|_\infty > (\log n)^{-\chi_1}\bigr) = \mathcal{O}\biggl(\frac{(d_n(\log n)^{\chi_1})^p}{n^{p/2}}\biggr), \qquad \chi_1 \geq 0.$$

LEMMA 4.2. *Assume that the assumptions of Theorem* 2.2 *are valid. Then we have*

$$P\Biggl(\Biggl\|n^{1/2}(\widehat{\mathbf{\Theta}}_{d_n} - \mathbf{\Theta}_{d_n}) - n^{-1/2}\sum_{k=1}^{n}\mathbf{U}_k\Biggr\|_\infty \geq (\log n)^{-\chi_1}\Biggr) = o(1),$$

where $1 < \chi_1$.

LEMMA 4.3. *Assume that the assumptions of Theorem* 2.2 *are valid. Then*:

(i) $$\lim_{n\to\infty} P\Biggl(\max_{1\leq h\leq d_n}\sigma^{-1}\Biggl|(n\gamma_{h,h}^{*})^{-1/2}\sum_{k=1}^{n}U_k^{(h)}\Biggr| \leq u_n\Biggr) = \exp(-\exp(-x)),$$

(ii) $$\sqrt{n}\|\widehat{\mathbf{\Phi}}_{d_n} - \mathbf{\Phi}_{d_n}\|_\infty = \mathcal{O}_P\bigl(\sqrt{\log d_n}\bigr),$$

where $u_n = a_n z + b_n$, a_n, b_n, z *are as in Theorem* 2.2.

The proofs of Lemmas 4.1, 4.2 and 4.3 are given in Section 5. Based on the above results, one readily derives the following weak version of Theorem 2.2.

COROLLARY 4.4. *Assume that the assumptions of Theorem 2.2 are valid. Then for $z \in \mathbb{R}$*

$$P\Big(a_n^{-1}\Big(\sqrt{n} \max_{1 \le h \le d_n} |(\gamma_{h,h}^* \sigma^2)^{-1/2}(\widehat{\theta}_i - \theta_i)| - b_n\Big) \le z\Big) \to \exp(-e^{-z}),$$

where a_n and b_n are as in Theorem 2.2.

COROLLARY 4.5. *Under the same conditions as in Theorem 2.2, we have*

$$|\widehat{\sigma}^2 - \sigma^2| = \mathcal{O}_P(n^{-1/2}\log n).$$

Throughout the proofs, the following inequality will be frequently used. For random variables $X_1, \dots, X_q$, and $\varepsilon > 0$, the inequality between the geometric and arithmetic mean implies

$$P\left(\prod_{i=1}^q |X_i| \ge \varepsilon\right) \le \sum_{i=1}^q P(|X_i| \ge \varepsilon^{1/q}). \tag{4.4}$$

PROOF OF COROLLARY 4.4. It holds that

$$\begin{aligned}&\max_{1 \le h \le d_n} \sigma^{-1}\left|(n\gamma_{h,h}^*)^{-1/2}\left(n(\widehat{\theta}_h - \theta_h) - \sum_{k=1}^n U_k^{(h)}\right)\right| \\ &\quad\le \sqrt{\Big(n\sigma^2 \inf_h \gamma_{h,h}^*\Big)^{-1}} \max_{1 \le h \le d_n}\left|\left(n(\widehat{\theta}_h - \theta_h) - \sum_{k=1}^n U_k^{(h)}\right)\right|.\end{aligned}$$

Since $\inf_h \gamma_{h,h}^* > 0$ and choosing $\chi_1 > 1$, the claim follows from Lemmas 4.2 and 4.3. □

PROOF OF COROLLARY 4.5. Trivially, it holds that

$$\begin{aligned}\widehat{\sigma}^2 - \sigma^2 &= \widehat{\phi}_0 - \phi_0 + \widehat{\mathbf{\Theta}}_{d_n}^T \widehat{\mathbf{\Phi}}_{d_n} - \mathbf{\Theta}_{d_n}^T \mathbf{\Phi}_{d_n} \\ &= \widehat{\phi}_0 - \phi_0 + (\widehat{\mathbf{\Theta}}_{d_n}^T - \mathbf{\Theta}_{d_n}^T)(\widehat{\mathbf{\Phi}}_{d_n} - \mathbf{\Phi}_{d_n}) \\ &\qquad + (\widehat{\mathbf{\Theta}}_{d_n}^T - \mathbf{\Theta}_{d_n}^T)\mathbf{\Phi}_{d_n} + \mathbf{\Theta}_{d_n}^T(\widehat{\mathbf{\Phi}}_{d_n} - \mathbf{\Phi}_{d_n}).\end{aligned}$$

By Corollary 4.4 and Lemma 4.3 we have

$$\begin{aligned}\|(\widehat{\mathbf{\Theta}}_{d_n}^T - \mathbf{\Theta}_{d_n}^T)(\widehat{\mathbf{\Phi}}_{d_n} - \mathbf{\Phi}_{d_n})\|_\infty &\le d_n \|\widehat{\mathbf{\Theta}}_{d_n}^T - \mathbf{\Theta}_{d_n}^T\|_\infty \|\widehat{\mathbf{\Phi}}_{d_n} - \mathbf{\Phi}_{d_n}\|_\infty \\ &= \mathcal{O}_P(n^{-1/2}\log n).\end{aligned}$$

Similarly, we obtain from Lemmas 4.3, 5.2 and Assumption 2.1

$$\begin{aligned}\|(\widehat{\mathbf{\Theta}}_{d_n}^T - \mathbf{\Theta}_{d_n}^T)\mathbf{\Phi}_{d_n}\|_\infty &= \mathcal{O}_P(n^{-1/2}\log n),\\ \|\mathbf{\Theta}_{d_n}^T(\widehat{\mathbf{\Phi}}_{d_n} - \mathbf{\Phi}_{d_n})\|_\infty &= \mathcal{O}_P(n^{-1/2}\log n).\end{aligned}$$

Moreover, from the above one readily deduces $|\widehat{\phi}_0 - \phi_0| = \mathcal{O}_P(n^{-1/2}\log n)$. Piecing everything together, the claim follows. □

PROOF OF THEOREM 2.2. Due to Corollary 4.4, it suffices to show that the error difference

$$\begin{aligned}\max_{1\leq i\leq d_n}\Delta_i &= \max_{1\leq i\leq d_n}\sqrt{n}|(\widehat{\gamma}_{i,i}^*\widehat{\sigma}^2)^{-1/2}(\widehat{\theta}_i - \theta_i) - (\gamma_{i,i}^*\sigma^2)^{-1/2}(\widehat{\theta}_i - \theta_i)|\\ &= \mathcal{O}_P((\log n)^{-\chi_1})\end{aligned}\tag{4.5}$$

for some $\chi_1 > 1$. Note that per assumption we have that $(\log n)^{\chi_2 p}n^{-p/2}d_n^p = \mathcal{o}(1)$ for some $\chi_2 > 1$. Moreover,

$$\begin{aligned}\max_{0\leq i\leq d_n}\Delta_i \leq \max_{0\leq i\leq d_n}&\big|((\widehat{\gamma}_{i,i}^*\widehat{\sigma}^2)^{1/2} - (\gamma_{i,i}^*\sigma^2)^{1/2})(\widehat{\gamma}_{i,i}^*\widehat{\sigma}^2)^{-1/2}\big|\\ &\times\sqrt{n}\max_{0\leq i\leq d_n}|(\widehat{\theta}_i - \theta_i)(\gamma_{i,i}^*\sigma^2)^{-1/2}|.\end{aligned}$$

Corollary 4.4 gives us $\sqrt{n}\max_{0\leq i\leq d_n}|(\widehat{\theta}_i - \theta_i)(\gamma_{i,i}^*\sigma^2)^{-1/2}| = \mathcal{O}_P(\log n)$, hence we need to study $|(\widehat{\gamma}_{i,i}^*\widehat{\sigma}^2)^{1/2} - (\gamma_{i,i}^*\sigma^2)^{1/2}|(\widehat{\gamma}_{i,i}^*\widehat{\sigma}^2)^{-1/2}$. Since

$$\begin{aligned}\left|\frac{(\widehat{\gamma}_{i,i}^*\widehat{\sigma}^2)^{1/2} - (\gamma_{i,i}^*\sigma^2)^{1/2}}{(\widehat{\gamma}_{i,i}^*\widehat{\sigma}^2)^{1/2}}\right| &= \left|\frac{\widehat{\gamma}_{i,i}^*\widehat{\sigma}^2 - \gamma_{i,i}^*\sigma^2}{(\widehat{\gamma}_{i,i}^*\widehat{\sigma}^2)^{1/2} + (\gamma_{i,i}^*\sigma^2)^{1/2}}\frac{1}{(\widehat{\gamma}_{i,i}^*\widehat{\sigma}^2)^{1/2}}\right|\\ &\leq \left|\frac{(\widehat{\gamma}_{i,i}^*\widehat{\sigma}^2 - \gamma_{i,i}^*\sigma^2)}{\widehat{\gamma}_{i,i}^*\widehat{\sigma}^2}\right|,\end{aligned}$$

it suffices to treat $(\widehat{\gamma}_{i,i}^*\widehat{\sigma}^2 - \gamma_{i,i}^*\sigma^2)(\widehat{\gamma}_{i,i}^*\widehat{\sigma}^2)^{-1}$. For $\varepsilon = \log n^{-\chi_2}$ we have

$$\begin{aligned}&\{|\sigma^2\gamma_{i,i}^* - \widehat{\sigma}^2\widehat{\gamma}_{i,i}^*| \geq \varepsilon\widehat{\gamma}_{i,i}^*\widehat{\sigma}^2\}\\ &\quad\subseteq \{|\sigma^2\gamma_{i,i}^* - \widehat{\sigma}^2\widehat{\gamma}_{i,i}^*|(1+\varepsilon) \geq \varepsilon\gamma_{i,i}^*\sigma^2\}\\ &\quad\subseteq \{|\sigma^2\gamma_{i,i}^* - \widehat{\sigma}^2\widehat{\gamma}_{i,i}^*| \geq \varepsilon\gamma_{i,i}^*\sigma^2/2\}\\ &\quad\subseteq \{|\sigma^2\widehat{\gamma}_{i,i}^* - \widehat{\sigma}^2\widehat{\gamma}_{i,i}^*| + |\sigma^2\gamma_{i,i}^* - \sigma^2\widehat{\gamma}_{i,i}^*| \geq \varepsilon\gamma_{i,i}^*\sigma^2/2\}.\end{aligned}$$

Since $\sigma^2, \gamma_{i,i}^* \geq C > 0$, we have from Lemma 4.1 that for $1 < \chi_1 < \chi_2$

$$P\Big(\max_{0\leq i\leq d_n}|\sigma^2\gamma_{i,i}^* - \sigma^2\widehat{\gamma}_{i,i}^*| \geq \log n^{-\chi_1}\min_{1\leq i\leq d_n}\gamma_{i,i}^*\Big) = \mathcal{O}(\log n^{\chi_2 p}n^{-p/2}d_n^p).$$

In order to treat $|\sigma^2\widehat{\gamma}_{i,i}^* - \widehat{\sigma}^2\widehat{\gamma}_{i,i}^*|$, note that

$$\max_{0\le i\le d_n} |\sigma^2\widehat{\gamma}_{i,i}^* - \widehat{\gamma}_{i,i}^*\widehat{\sigma}^2| \le \max_{0\le i\le d_n} (|\sigma^2 - \widehat{\sigma}^2||\gamma_{i,i}^* - \widehat{\gamma}_{i,i}^*| + \gamma_{i,i}^*|\sigma^2 - \widehat{\sigma}^2|),$$

which by virtue of Corollary 4.5 and Lemma 4.1 is of the magnitude $\mathcal{O}_P(n^{-1/2} \times \log n)$. We thus obtain that

$$P\Bigl(\max_{0\le i\le d_n} \Delta_i \ge \log n^{-\chi_1}\Bigr) = \mathcal{o}(1)$$

for some $\chi_1 > 1$, which completes the proof. □

PROOF OF COROLLARY 2.5. First note that both conditions (i) and (ii) imply that $|\alpha_i| = \mathcal{O}(\rho^{-i})$, $0 < \rho < 1$ (cf. [17]). Hence Remark 2.3 yields that we may choose $d_n = \mathcal{O}(n^\delta)$, $0 < \delta < 1/2$. Now assume that (i) holds. Then relation (4.1) implies

$$\sigma^2 \inf_h \gamma_{h,h}^* \ge 1 - \sum_{i=1}^{d_n} |\theta_i|^2 \ge 1 - \sum_{i=1}^{d_n} |\theta_i| > 0,$$

whence the claim. If (ii) holds, then for large enough n we obtain similarly

$$\sigma^2 \inf_h \gamma_{h,h}^* \ge \sum_{i=0}^{\alpha} \theta_i^2 - \sum_{i=\beta}^{d_n} \theta_i^2 \ge \sum_{i=0}^{\alpha} \theta_i^2 \ge 1,$$

where α, β are as in (4.1). □

We are now ready to prove Theorem 2.8.

PROOF OF THEOREM 2.8. Let $q_0 = q$ be the true order of the AR(q)-process $\{X_k\}_{k\in\mathbb{Z}}$, put

$$\overline{\theta}_{i,n} = a_n^{-1}\bigl(\sqrt{n}|(\widehat{\gamma}_{i,i}^*\widehat{\sigma}^2)^{-1/2}(\widehat{\theta}_i - \theta_i)| - b_n\bigr)$$

and assume first that $k \in \mathbb{N}$, $k > 0$. Note that $\theta_i = 0$ for $i > q$. Then we have that

$$\begin{aligned} P(\widehat{q}_{z_n} = k + q) &= P\Bigl(\{\overline{\theta}_{q+k,n} > z_n\} \cap \Bigl\{\max_{k+q+1\le i\le d_n} \overline{\theta}_{i,n} \le z_n\Bigr\}\Bigr) \\ &= P\Bigl(\max_{k+q\le i\le d_n} \overline{\theta}_{i,n} \le z_n\Bigr) - P\Bigl(\max_{k+q+1\le i\le d_n} \overline{\theta}_{i,n} \le z_n\Bigr). \end{aligned}$$

Due to Theorem 6.1, we can approximate the sequence $\{\overline{\theta}_{i,n}\}_{1\le i\le d_n}$ by a suitably transformed corresponding sequence of mean-zero Gaussian random variables $\boldsymbol{\xi}_{d_n} = (\xi_{n,1}, \dots, \xi_{n,d_n})^T$ with covariance matrix $\boldsymbol{\Gamma}_{\boldsymbol{\xi}_{d_n}}^*$. Let $\boldsymbol{\eta}_{d_n} = (\eta_{n,1}, \dots, \eta_{n,d_n})^T$ be another sequence of i.i.d. mean-zero Gaussian random variables with unit variance. Following Deo [18], we obtain from $\max|\boldsymbol{\Gamma}_{\boldsymbol{\xi}_{d_n}}^* - \boldsymbol{\Gamma}_{d_n}^*| =$

$\mathcal{O}(d_n^{-1})$ that for fixed $l \in \mathbb{N}$

$$\Big| P\Big(\max_{q+l \le i \le d_n} a_n^{-1}(|\xi_{n,i}| - b_n) \le z_n \Big) - P\Big(\max_{q+l \le i \le d_n} a_n^{-1}(|\eta_{n,i}| - b_n) \le z_n \Big) \Big|$$

$$\le C \sum_{1 \le i < j \le d_n} |\rho_{i,j}| \big(d_n^{-2z_n^2/(1+|\rho_{i,j}|)} \big).$$

Imitating the technique in Berman [9], we obtain that the above quantity is of the magnitude $\mathcal{O}(d_n^{(-z_n^2+1)/2})$. This yields

$$\begin{aligned} P(\widehat{q}_{z_n} = k+q) &\\ = P\Big(&\max_{q+k+1 \le i \le d_n} a_n^{-1}(|\eta_{n,i}| - b_n) \le z_n \Big) \\ &- P\Big(\max_{q+k \le i \le d_n} a_n^{-1}(|\eta_{n,i}| - b_n) \le z_n \Big) + \mathcal{O}\big(n^{-\nu} + d_n^{(-z_n^2+1)/2}\big) \\ = P\big(&a_n^{-1}(|\eta_{n,1}| - b_n) \le z_n\big)^{d_n-k-q} \big(1 - P\big(a_n^{-1}(|\eta_{n,1}| - b_n) \le z_n\big)\big) \\ &+ \mathcal{O}\big(n^{-\nu} + d_n^{(-z_n^2+1)/2}\big). \end{aligned}$$

From the definition of a_n, b_n, and since $z_n \to \infty$, we obtain that (Deo [18])

$$\lim_n P\big(a_n^{-1}(|\eta_{n,1}| - b_n) \le z_n\big)^{d_n-k-q} \to 1, \tag{4.6}$$

$$P\big(a_n^{-1}(|\eta_{n,1}| - b_n) > z_n\big) = \frac{e^{-z_n}}{d_n} + \mathcal{O}\Big(\frac{e^{-z_n}}{d_n}\Big). \tag{4.7}$$

This yields

$$P(\widehat{q}_{z_n} = k+q) = \frac{e^{-z_n}}{d_n} + \mathcal{O}\Big(\frac{e^{-z_n}}{d_n} + d_n^{(-z_n^2+1)/2}\Big) \tag{4.8}$$

and in particular

$$P(\widehat{q}_{z_n} > q) = \sum_{k=1}^{d_n} P(\widehat{q}_{z_n} = k+q) = e^{-z_n} + \mathcal{O}(e^{-z_n} + d_n^{-z_n^2+2}), \tag{4.9}$$

and per assumption the right-hand side goes to zero as n increases. We now consider the case $P(\widehat{q}_{z_n} < q)$. To this end, let $k \in \mathbb{N}$, $k > 0$. Then we have

$$\begin{aligned} P(\widehat{q}_{z_n} = q-k) \le P(\overline{\theta}_{q-k,n} \le z_n) &\\ = P\big(a_n^{-1}\big(&|\xi_{n,q-k} + \sqrt{n}\theta_{q-k}| - b_n\big) \le z_n\big) \\ &+ \mathcal{O}(n^{-\nu}). \end{aligned}$$

Since $|\theta_{q-k}| > 0$, one readily verifies by known properties of the Gaussian c.d.f. that $P(a_n^{-1}(|\xi_{n,q-k} + \sqrt{n}\theta_{q-k}| - b_n) \le z_n) = \mathcal{O}(n^{-\nu})$, and hence

$$P(\widehat{q}_{z_n} = q-k) = \mathcal{O}(n^{-\nu}) \tag{4.10}$$

and in particular

$$P(\widehat{q}_{z_n} < q) = \mathcal{O}(d_n n^{-\nu}) \to 0 \tag{4.11}$$

as n increases. This together with (4.9) establishes consistency. □

PROOF OF THEOREM 2.12. Let $q_0 = q$ be the true order of the AR(q)-process $\{X_k\}_{k\in\mathbb{Z}}$. The proof then consists of two parts. It is first shown that $P(\widehat{q}_n^* < q) \to 0$, whereas in the second part the claim $P(\widehat{q}_n^* > q) \to 0$ is established.

First note that Lemma 5.2 and the Cauchy interlacing theorem yield that $\|\mathbf{\Gamma}_k\|_\infty$, $\|\mathbf{\Gamma}_k^{-1}\|_\infty \le C < \infty$, uniformly for $1 \le k \le d_n$. Hence, using that $\widehat{\mathbf{\Gamma}}_k^{-1}\widehat{\mathbf{\Phi}}_k = \widehat{\mathbf{\Theta}}_k$, Lemma 4.3 and a slight adaption of Lemma 4.1 imply that

$$|\widehat{\sigma}^2(k) - \sigma^2(k)| = o_P(1) \qquad \text{uniformly for } 1 \le k \le q.$$

Since $\inf_h |1 - \theta_h^2| > 0$, we conclude that $\inf_k \sigma^2(k) > 0$ and hence

$$|\log(\widehat{\sigma}^2(k)) - \log(\sigma^2(k))| = o_p(1). \tag{4.12}$$

By Hannan [23], Chapter VI, it holds that for $k \in \mathbb{N}$

$$\log(\widehat{\sigma}^2(k)) = \log\widehat{\phi}_{n,0} + \sum_{j=1}^{k} \log\bigl(1 - \widehat{\theta}_j^2(k)\bigr). \tag{4.13}$$

Then, arguing as in Hannan and Quinn [25], we have due to $C_n = o(n)$ that for large enough n

$$f_n(k) = \log(\widehat{\sigma}^2(k)) + n^{-1}C_n k$$

is a decreasing function in k for $0 \le k < q$, and strictly decreasing for $q - 1 \le k \le q$ (since $\theta_q^2 > 0$) with probability approaching one. This implies that eventually $\widehat{q}_n^* \ge q$, hence it suffices to establish that the probability of overestimating the order goes to zero as n increases, that is,

$$\lim_n P\Bigl(\arg\min_{q\le k\le d_n}\bigl(\log(\widehat{\sigma}^2(k)) + n^{-1}C_n k\bigr) \ge q + 1\Bigr) = 0. \tag{4.14}$$

Using the same arguments as in [3], it follows that it suffices to establish

$$\lim_n P\left(\max_{1\le k\le d_n - q}\left(\sum_{j=1+q}^{k+q} -\log\bigl(1 - \widehat{\theta}_j^2(k)\bigr) - n^{-1}C_n k\right) \ge 0\right) = 0. \tag{4.15}$$

By Theorem 2.2, we have that

$$\|\widehat{\mathbf{\Theta}}_k^2\|_\infty = \mathcal{O}_P(n^{-1}\log d_n) \qquad \text{for } q_0 < k \le d_n. \tag{4.16}$$

This implies that for some increasing $\chi_n \to \infty$, we obtain that

$$-\sum_{j=1+q}^{k+q} \log\bigl(1 - \widehat{\theta}_j^2(k)\bigr) \le k\chi_n n^{-1}\log d_n, \tag{4.17}$$

with probability approaching one. Since $\log d_n = o(C_n)$ per assumption, (4.15) follows, which completes the proof. □

5. Proofs of the auxiliary results of Section 4.

The following result is required for the proofs.

LEMMA 5.1. *Let $\{X_k\}_{k\in\mathbb{Z}}$ be an* AR(q) *process such that Assumption* 2.1 *is satisfied. Then*:

(i) $\sum_{h=0}^{\infty}|\mathrm{Cov}(X_k, X_{k+h})| < \infty$,

(ii) $\sqrt{n}\|\widehat{\phi}_{n,h} - \phi_h\|_p = \mathcal{O}(1)$, $p \geq 1$.

PROOF. Both properties (i), (ii) follow from Assumption 2.1 via straightforward computations (cf. [17, 23]). □

Recall the notation $\boldsymbol{\Gamma}_m = (\gamma_{i,j})_{1\leq i,j\leq m}$ and $\boldsymbol{\Gamma}_m^{-1} = (\gamma_{i,j}^*)_{1\leq i,j\leq m}$ for the covariance matrix and its inverse.

LEMMA 5.2. *Assume that Assumption* 2.1 *holds. Then for $d_n \leq m$ we have $\|\boldsymbol{\Gamma}_m\|_\infty$, $\|\boldsymbol{\Gamma}_m^{-1}\|_\infty \leq C < \infty$, uniformly in m.*

PROOF. Using relation (4.1) and the corresponding notation, one obtains

$$\begin{aligned}\|\boldsymbol{\Gamma}_m^{-1}\|_\infty &= \sigma^{-2}\max_{1\leq j\leq m}\sum_{i=1}^{m}|\sigma^2\gamma_{i,j}^*| \\ &\leq 2\sigma^{-2}\max_{1\leq j\leq m}\left|\sum_{r=0}^{\alpha}\theta_r\theta_{r+j-i} - \sum_{r=\beta}^{d_n+i-j}\theta_r\theta_{r+j-i}\right| \\ &\leq 4\sigma^{-2}\sum_{|h|\leq m}\sum_{r=0}^{m}|\theta_r\theta_{|r+h|}| \leq 8\sigma^{-2}\left(\sum_{r=0}^{\infty}|\theta_r|\right)^2,\end{aligned}$$

where $\theta_h = 0$ for $h < 0$. Due to Assumption 2.1, the above expression is finite, hence the first claim follows. In order to establish the result for $\boldsymbol{\Gamma}_m$, note that

$$\|\boldsymbol{\Gamma}_m\|_\infty = \max_{1\leq j\leq m}\sum_{i=1}^{m}|\gamma_{i,j}| \leq 2\sum_{h=0}^{\infty}|\phi_h| < \infty$$

by Lemma 5.1(i), which yields the claim. □

We can now prove Lemma 4.3, which we reformulate below for the sake of readability.

LEMMA 5.3. *Suppose that* $\inf_h|\gamma_{h,h}^*| > 0$ *and Assumption* 2.1 *holds. Then*:

(i) $$\lim_{n\to\infty}P\left(\max_{0\leq h\leq d_n}\sigma^{-1}\left|(n\gamma_{h,h}^*)^{-1/2}\sum_{k=1}^{n}U_k^{(h)}\right| \leq u_n\right) = \exp(-\exp(-x)),$$

(ii) $$\sqrt{n}\|\widehat{\boldsymbol{\Phi}}_{d_n} - \boldsymbol{\Phi}_{d_n}\|_\infty = \mathcal{O}_P\big(\sqrt{\log d_n}\big).$$

PROOF. We will first show (i). Using the notation established in Section 4, we have

$$U_k^{(h)} = \varepsilon_k \left(\sum_{j=1}^{d_n} \gamma_{h,j}^{(*)} \sum_{i=0}^{\infty} \alpha_i \varepsilon_{k-j-i} \right) := \varepsilon_k \sum_{r=1}^{\infty} \alpha_{r,h}^{*} \varepsilon_{k-r},$$

where $\alpha_{r,h}^{*} = \sum_{\{i \geq 0, j \geq 0, i+j=r\}} \gamma_{h,j}^{*} \alpha_i$. Let $0 < \delta < \delta^*$, and put $m_n = \lfloor n^{\delta^*} \rfloor$. Then it follows from Lemma 5.2 that

$$\sup_h \sum_{r=m_n}^{\infty} |\alpha_{r,h}^{*}| \leq C \sum_{i=m_n-d_n}^{\infty} |\alpha_i| = \mathcal{O}\bigl((m_n - d_n)^{-\vartheta}\bigr) = \mathcal{O}(m_n^{-\vartheta}).$$

Due to Assumption 2.1, one may thus repeat the (quite lengthy) proof of Theorem 1 (see also Remark 2) in [48] to obtain the result. In fact, the present case is easier to handle, since $\{U_k^{(h)}\}_{k \in \mathbb{N}}$ is a martingale sequence.

Assertion (ii) follows directly from Theorem 1 in [48]. □

We can now proof Lemma 4.1, which we restate for the sake of readability.

LEMMA 5.4. *If Assumption 2.1 holds, we have for* $\chi_1 > 0$

$$P\bigl(\|\widehat{\mathbf{\Gamma}}_{d_n}^{-1} - \mathbf{\Gamma}_{d_n}^{-1}\|_\infty \geq (\log n)^{-\chi_1}\bigr) = \mathcal{O}\biggl(\frac{(d_n (\log n)^{\chi_1})^p}{n^{p/2}}\biggr).$$

PROOF. We introduce the following abbreviations. Put

$$E = \|\mathbf{\Gamma}_{d_n}^{-1}\|_\infty, \qquad F = \|\widehat{\mathbf{\Gamma}}_{d_n}^{-1} - \mathbf{\Gamma}_{d_n}^{-1}\|_\infty, \qquad G = \|\widehat{\mathbf{\Gamma}}_{d_n} - \mathbf{\Gamma}_{d_n}\|_\infty.$$

Due to the stationarity of $\{X_k\}_{k \in \mathbb{Z}}$ it follows that

$$G = \|\widehat{\mathbf{\Gamma}}_{d_n} - \mathbf{\Gamma}_{d_n}\|_\infty \leq 2 \sum_{h \leq d_n} \bigl|\widehat{\phi}_{n,|h|} - \phi_{|h|}\bigr|, \tag{5.1}$$

and thus an application of the Hölder and Minikowski inequalities yields

$$\mathbb{E}(|G|) \leq 2 \sum_{h \leq d_n} \bigl\|\widehat{\phi}_{n,|h|} - \phi_{|h|}\bigr\|_p. \tag{5.2}$$

Due to Lemma 5.1(ii) we have $\sqrt{n}\|\widehat{\phi}_{n,|i-j|} - \phi_{|i-j|}\|_p \leq C_p$ for some finite constant C_p, thus the Markov inequality in connection with Minikowski's inequality implies

$$P\bigl(\|\widehat{\mathbf{\Gamma}}_{d_n} - \mathbf{\Gamma}_{d_n}\|_\infty \geq (\log n)^{-\chi_1}\bigr) = \mathcal{O}\biggl(\frac{(d_n (\log n)^{\chi_1})^p}{n^{p/2}}\biggr). \tag{5.3}$$

Due to the sub-multiplicativity of the matrix norm $\|\cdot\|_\infty$, proceeding as in Lemma 3 in [7] one obtains

$$F \leq (E + F) G E,$$

and in particular if $EG < 1$

$$F \leq E^2 G/(1 - EG).$$

Since we have $E < \infty$ due to Lemma 5.2, we deduce that for sufficiently large n

$$P(F \geq \varepsilon) \leq P\big(G \geq (\log n)^{-1}\big) + P(G \geq E^2/2\varepsilon).$$

Choosing $\varepsilon = (\log n)^{-\chi_1}$, the claim follows. □

We are now in the position to show Lemma 4.1. Recall that we have

$$\mathbf{Y} = \mathbf{X}\mathbf{\Phi}_d + \mathbf{Z}, \tag{5.4}$$

where $\mathbf{Y} = (X_1, \dots, X_n)^T$, $\mathbf{Z} = (\varepsilon_1, \dots, \varepsilon_n)^T$, and $\mathbf{X}$ is the $n \times d_n$ design matrix.

We introduce the estimator $\widetilde{\mathbf{\Theta}} = (\widetilde{\theta}_1, \dots, \widetilde{\theta}_d)^T$ via

$$\widetilde{\mathbf{\Theta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}. \tag{5.5}$$

REMARK 5.5. It is evident from the proof that Lemma 5.4 remains valid if one replaces $\widehat{\mathbf{\Gamma}}_{d_n}$ with $n(\mathbf{X}^T\mathbf{X})^{-1}$, which in fact is the better estimator.

PROPOSITION 5.6. *Let $\{X_k\}_{k\in\mathbb{Z}}$ be an* AR(d_n) *process, such that the assumptions of Theorem 2.2 are satisfied. Then*

$$P\big(\|\sqrt{n}(\widehat{\mathbf{\Theta}} - \widetilde{\mathbf{\Theta}})\|_\infty \geq (\log n)^{-\chi_1}\big) = \mathcal{O}((\log n)^{\chi_1 p/2} n^{-p/4} d_n^{p/4+1}) + \mathcal{o}(1).$$

PROOF. Following the proof of [17], Theorem 8.10.1, we have the following decomposition:

$$\sqrt{n}(\widehat{\mathbf{\Theta}} - \widetilde{\mathbf{\Theta}}) = \sqrt{n}\widehat{\mathbf{\Gamma}}_{d_n}^{-1}(\widehat{\mathbf{\Phi}}_{d_n} - n^{-1}\mathbf{X}^T\mathbf{Y}) + n^{1/2}(\widehat{\mathbf{\Gamma}}_{d_n}^{-1} - n(\mathbf{X}^T\mathbf{X})^{-1})n^{-1}\mathbf{X}^T\mathbf{Y}.$$

For the ith component of $\sqrt{n}(\widehat{\mathbf{\Phi}}_{d_n} - n^{-1}\mathbf{X}^T\mathbf{Y})$, which we denote with Υ_i, we have

$$n^{-1/2}\sum_{k=1-i}^{0} X_k X_{k+i} + \sqrt{n}\overline{X}_n\left((1 - n^{-1}i)\overline{X}_n - n^{-1}\sum_{k=1}^{n-i}(X_k + X_{k+i})\right).$$

Using the Minikowski and the Cauchy–Schwarz inequalities we get

$$\begin{aligned}
&\left\| n^{-1/2}\sum_{k=1-i}^{0} X_k X_{k+i} + \sqrt{n}\overline{X}_n\left((1 - n^{-1}i)\overline{X}_n - n^{-1}\sum_{k=1}^{n-i}(X_k + X_{k+i})\right)\right\|_{p/2} \\
&\quad \leq \sqrt{\frac{|1-i|}{n}}\left\| |1-i|^{-1/2}\sum_{k=1-i}^{0}(X_k X_{k+i} - \phi_i)\right\|_{p/2} + n^{-1/2}\sum_{k=1-i}^{0}|\phi_i| \\
&\qquad + \|\sqrt{n}\overline{X}_n\|_p\left(\|\overline{X}_n\|_p + n^{-1/2}\left\| n^{-1/2}\sum_{k=1}^{n-i}(X_k + X_{k+i})\right\|_p\right) \\
&\quad := A_n.
\end{aligned}$$

Since $0 \le i \le d_n$, we obtain from Lemma 5.1 that $A_n = \mathcal{O}(n^{-1/2} d_n^{1/2})$, and hence by the Markov inequality

$$\begin{aligned} P(\|\sqrt{n}(\widehat{\mathbf{\Phi}}_{d_n} - n^{-1}\mathbf{X}^T\mathbf{Y})\|_\infty \ge \varepsilon) \\ \le \sum_{i=1}^{d_n} P(|\Upsilon_i| \ge \varepsilon) = \mathcal{O}(\varepsilon^{-p/2} n^{-p/4} d_n^{p/4+1}). \end{aligned} \tag{5.6}$$

Put $B_n = \widehat{\mathbf{\Phi}}_{d_n} - n^{-1}\mathbf{X}^T\mathbf{Y}$. Then by adding and subtracting $\mathbf{\Gamma}_{d_n}^{-1}$ we obtain

$$\begin{aligned} P(\sqrt{n}\|\widehat{\mathbf{\Gamma}}_{d_n}^{-1} B_n\|_\infty \ge \varepsilon) \le P(\sqrt{n}\|(\widehat{\mathbf{\Gamma}}_{d_n}^{-1} - \mathbf{\Gamma}_{d_n}^{-1})B_n\|_\infty \ge \varepsilon/2) \\ + P(\sqrt{n}\|\mathbf{\Gamma}_{d_n}^{-1} B_n\|_\infty \ge \varepsilon/2). \end{aligned} \tag{5.7}$$

In order to control the first expression, note that

$$\begin{aligned} P(\sqrt{n}\|(\widehat{\mathbf{\Gamma}}_{d_n}^{-1} - \mathbf{\Gamma}_{d_n}^{-1})B_n\|_\infty \ge \varepsilon/2) \le P(\|\widehat{\mathbf{\Gamma}}_{d_n}^{-1} - \mathbf{\Gamma}_{d_n}^{-1}\|_\infty \varepsilon \ge \varepsilon/2) \\ + P(\|B_n\|_\infty \ge \varepsilon), \end{aligned}$$

which by Lemma 5.4 and (5.6) is of the magnitude $\mathcal{O}(\varepsilon^{-p/2} n^{-p/4} d_n^{p/4+1})$. Moreover, since $\|\mathbf{\Gamma}_{d_n}^{-1}\|_\infty < \infty$ by Lemma 5.2, the bound in (5.6) implies that for some $C > 0$

$$P(\sqrt{n}\|\mathbf{\Gamma}_{d_n}^{-1} B_n\|_\infty \ge \varepsilon/2) \le P(\sqrt{n}\|B_n\|_\infty \ge \varepsilon C^{-1}) = \mathcal{O}(\varepsilon^{-p/2} n^{-p/4} d_n^{p/4+1}),$$

hence we conclude that

$$P(\|\sqrt{n}\widehat{\mathbf{\Gamma}}_{d_n}^{-1}(\widehat{\mathbf{\Phi}}_{d_n} - n^{-1}\mathbf{X}^T\mathbf{Y})\|_\infty \ge \varepsilon) = \mathcal{O}(\varepsilon^{-p/2} n^{-p/4} d_n^{p/4+1}). \tag{5.8}$$

We will now treat the second part, which we rewrite as

$$\begin{aligned} n^{1/2}(\widehat{\mathbf{\Gamma}}_{d_n}^{-1} - n(\mathbf{X}^T\mathbf{X})^{-1})(n^{-1}\mathbf{X}^T\mathbf{Y} - n^{-1}\mathbb{E}(\mathbf{X}^T\mathbf{Y})) \\ + n^{1/2}(\widehat{\mathbf{\Gamma}}_{d_n}^{-1} - n(\mathbf{X}^T\mathbf{X})^{-1})n^{-1}\mathbb{E}(\mathbf{X}^T\mathbf{Y}) \\ =: C_n + D_n. \end{aligned}$$

Due to Lemma 5.3 (requires an easy adaption), we have

$$\|n^{-1/2}\mathbf{X}^T\mathbf{Y} - n^{-1/2}\mathbb{E}(\mathbf{X}^T\mathbf{Y})\|_\infty = \mathcal{O}_P(\log n). \tag{5.9}$$

Moreover, it holds that

$$\sqrt{n}(\widehat{\mathbf{\Gamma}}_{d_n}^{-1} - n(\mathbf{X}^T\mathbf{X})^{-1}) = \widehat{\mathbf{\Gamma}}_{d_n}^{-1}\sqrt{n}(n^{-1}(\mathbf{X}^T\mathbf{X}) - \widehat{\mathbf{\Gamma}}_{d_n})n(\mathbf{X}^T\mathbf{X})^{-1},$$

and thus the sub-multiplicativity of the matrix norm $\|\cdot\|_\infty$ implies

$$\begin{aligned} \|C_n\|_\infty \le \|\widehat{\mathbf{\Gamma}}_{d_n}^{-1}\|_\infty \|\sqrt{n}(n^{-1}(\mathbf{X}^T\mathbf{X}) - \widehat{\mathbf{\Gamma}}_{d_n})\|_\infty \|n(\mathbf{X}^T\mathbf{X})^{-1}\|_\infty \\ \times \|n^{-1/2}\mathbf{X}^T\mathbf{Y} - n^{-1/2}\mathbb{E}(\mathbf{X}^T\mathbf{Y})\|_\infty. \end{aligned}$$

Using (5.9) we thus obtain

$$\begin{aligned}&P(\|C_n\|_\infty \geq \varepsilon)\\&\quad\leq \mathcal{O}(1) + P\big(\|\widehat{\mathbf{\Gamma}}_{d_n}^{-1}\|_\infty \|\sqrt{n}\big(n^{-1}(\mathbf{X}^T\mathbf{X}) - \widehat{\mathbf{\Gamma}}_{d_n}\big)\|_\infty\\&\qquad\qquad\qquad\qquad \times \|n(\mathbf{X}^T\mathbf{X})^{-1}\|_\infty \log n \geq \varepsilon\big).\end{aligned}$$

Put $\Delta_n = n^{-1}(\mathbf{X}^T\mathbf{X}) - \widehat{\mathbf{\Gamma}}_{d_n}$. By adding and subtracting $\mathbf{\Gamma}_{d_n}^{-1}$ and using Lemma 5.4 (see Remark 5.5) and Lemma 5.2 we obtain

$$\begin{aligned}&P\big(\|\widehat{\mathbf{\Gamma}}_{d_n}^{-1}\|_\infty \|\Delta_n\|_\infty \|n(\mathbf{X}^T\mathbf{X})^{-1}\|_\infty \log n \geq \varepsilon\big)\\&\quad\leq 2P(\|\Delta_n\|_\infty \log n \geq 1) + P(\|\mathbf{\Gamma}_{d_n}^{-1} - \widehat{\mathbf{\Gamma}}_{d_n}^{-1}\|_\infty \geq \varepsilon)\\&\qquad + P\big(\|\mathbf{\Gamma}_{d_n}^{-1} - n^{-1}(\mathbf{X}^T\mathbf{X})\|_\infty \geq \varepsilon\big).\end{aligned}$$

Choosing $\varepsilon = (\log n)^{-\chi_1}$, Lemma 5.4 and (5.6) thus yield the bound

$$P\big(\|\widehat{\mathbf{\Gamma}}_{d_n}^{-1}\|_\infty \|\Delta_n\|_\infty \log n \geq (\log n)^{-\chi_1}\big) = \mathcal{O}((\log n)^{\chi_1 p} n^{-p/2} d_n^{p/2}). \tag{5.10}$$

Piecing everything together, the claim follows. □

We are now in the position to proof Lemma 4.2.

PROOF OF LEMMA 4.2. We have that

$$\begin{aligned}&P\big(\|n^{1/2}(\widehat{\mathbf{\Theta}} - \mathbf{\Theta}) - n^{-1/2}\mathbf{\Gamma}^{-1}\mathbf{X}^T\mathbf{Z}\|_\infty \geq 2\varepsilon\big)\\&\quad\leq P\big(\|n^{1/2}(\widehat{\mathbf{\Theta}} - \widetilde{\mathbf{\Theta}})\|_\infty \geq \varepsilon\big) + P\big(\|n^{1/2}(\widetilde{\mathbf{\Theta}} - \mathbf{\Theta}) - n^{-1/2}\mathbf{\Gamma}^{-1}\mathbf{X}^T\mathbf{Z}\|_\infty \geq \varepsilon\big).\end{aligned}$$

Setting $\varepsilon = \log n^{-\chi_1}$, $\chi_1 > 2$, Proposition 5.6 implies that

$$\|n^{1/2}(\widehat{\mathbf{\Theta}} - \widetilde{\mathbf{\Theta}})\|_\infty = \mathcal{O}_P(\log n^{-\chi_1}).$$

Moreover, the proof of Proposition 5.6 gives us

$$n^{1/2}(\widetilde{\mathbf{\Theta}} - \mathbf{\Theta}) - n^{-1/2}\mathbf{\Gamma}^{-1}\mathbf{X}^T\mathbf{Z} = \big(n(\mathbf{X}^T\mathbf{X})^{-1} - \mathbf{\Gamma}^{-1}\big)n^{-1/2}\mathbf{X}^T\mathbf{Z}, \tag{5.11}$$

and hence Remark 5.5 and Lemma 5.3 imply that

$$\begin{aligned}&\|n^{1/2}(\widetilde{\mathbf{\Theta}} - \mathbf{\Theta}) - n^{-1/2}\mathbf{\Gamma}^{-1}\mathbf{X}^T\mathbf{Z}\|_\infty\\&\quad\leq \|n(\mathbf{X}^T\mathbf{X})^{-1} - \mathbf{\Gamma}^{-1}\|_\infty \|n^{-1/2}\mathbf{X}^T\mathbf{Z}\|_\infty\\&\quad= \mathcal{O}_P(\log n^{-\chi_1+1}),\end{aligned} \tag{5.12}$$

which completes the proof. □

6. Gaussian approximation. In this section we obtain, under suitable assumptions, a normal approximation for the quantity $n^{-1/2}\mathbf{\Gamma}^{-1}\mathbf{X}^T\mathbf{Z}$, where we use the notation introduced in Section 4. This entitles us to obtain a quantitative version of Theorem 2.2 under stronger conditions. Let $\mathbf{V}_k = (X_{k-1}, \dots, X_{k-d_n})^T \varepsilon_k$, $k \in \mathbb{N}$. We have

$$n^{-1/2}\mathbf{\Gamma}^{-1}\mathbf{X}^T\mathbf{Z} = n^{-1/2}\mathbf{\Gamma}^{-1}\sum_{k=1}^{n}\mathbf{V}_k = n^{-1/2}\sum_{k=1}^{n}\mathbf{U}_k,$$

where $\mathbf{V}_k = (V_k^{(1)}, \dots, V_k^{(d_n)})^T$, $\mathbf{U}_k = (U_k^{(1)}, \dots, U_k^{(d_n)})^T$. Note that $\mathbf{V}_k$ and $\mathbf{U}_k$ are both martingale sequences. In particular, it holds that $\mathbb{E}(\mathbf{V}_k) = \mathbb{E}(\mathbf{U}_k) = 0$ and

$$\mathbb{E}(\mathbf{V}_k\mathbf{V}_{k+h}^T) = \begin{cases} \sigma^2\Gamma_{d_n}, & \text{if } h = 0, \\ 0_{d_n \times d_n}, & \text{if } h \neq 0, \end{cases} \tag{6.1}$$

since $\boldsymbol{\varepsilon}_k$ is independent of $\{X_{k-i}\}_{i \geq 1}$. Throughout this section, we will always assume that $d_n = \mathcal{O}(n)$.

The main theorem is formulated below.

THEOREM 6.1. *Suppose that Assumption 2.7 holds. If $d_n = \mathcal{O}(n^\delta)$ with $\delta < 1/7$, then on a possible larger probability space, there exists a d_n-dimensional Gaussian random vector $\mathbf{Z}$ with covariance matrix $\mathbf{\Gamma}_Z$, such that*

$$P\left(\left\|\mathbf{Z} - \sum_{k=1}^{n}\mathbf{U}_k\right\|_\infty \geq v_n\right) = \mathcal{O}(n^{-\nu}),$$

where $v_n = \sqrt{n}(\log n)^{-\chi_3}$, for arbitrary ν, $\chi_3 \geq 0$, and $\max \|n^{-1}\mathbf{\Gamma}_Z - \sigma^2\mathbf{\Gamma}_{d_n}\| = \mathcal{O}(d_n^{-1})$.

REMARK 6.2. If one succeeds in establishing a quantitative version of Lemma 21 in [48] with an appropriate error bound, corresponding results to Theorem 6.1 with $0 < \delta < 1$ should be possible. This, however, is beyond the scope of the present paper.

The proof of Theorem 6.1 partially follows [8], Theorem 4.1, and is based on a series of lemmas. To this end, we require some preliminary notation. For a d-dimensional vector $\mathbf{x} = (x_1, \dots, x_d)$, we denote with $|\mathbf{x}|_d = (\sum_{i=1}^n (x_i)^2)^{1/2}$ the usual Euclidean norm. The following coupling inequality is due to Berthet and Mason [10].

LEMMA 6.3 (Coupling inequality). *Let $X_1, \dots, X_N$ be independent, mean-zero random vectors in $\mathbb{R}^n$, $n \geq 1$, such that for some $B > 0$, $|X_i|_n \leq B$, $i = 1, \dots, N$. If the probability space is rich enough, then for each $\delta > 0$, one can define independent normally distributed mean-zero random vectors $\xi_1, \dots, \xi_N$ with*

ξ_i and X_i having the same variance/covariance matrix for $i = 1, \dots, N$, such that for universal constants $C_1 > 0$ and $C_2 > 0$,

$$P\left\{\left|\sum_{i=1}^{N}(X_i - \xi_i)\right|_n > \delta\right\} \le C_1 n^2 \exp\left(-\frac{C_2\delta}{Bn^2}\right).$$

The proof of Theorem 6.1 is based on a blocking argument, which in turn requires carefully truncated random variables. Put

$$n^{-1/2}\mathbf{\Gamma}^{-1}\mathbf{X}^T\mathbf{Z} = n^{-1/2}\mathbf{\Gamma}^{-1}\sum_{k=1}^{n}\mathbf{V}_k = n^{-1/2}\sum_{k=1}^{n}\mathbf{U}_k,$$

where $\mathbf{U}_k = (U_k^{(1)}, \dots, U_k^{(d_n)})^T$. Note that $\mathbf{V}_k$ and $\mathbf{U}_k$ are both martingale sequences.

LEMMA 6.4. *Suppose that Assumption 2.7 holds. Then for $q \ge 3$:*

$$\text{(i)} \quad P\left(\left\|n^{-1/2}\sum_{k=1}^{n}\mathbf{U}_k\right\|_\infty \ge \sqrt{q\log n}\right) = \mathcal{O}(n^{-\nu}),$$

$$\text{(ii)} \quad P\left(\sqrt{n}\|\widehat{\mathbf{\Phi}}_{d_n} - \mathbf{\Phi}_{d_n}\|_\infty \ge \sqrt{q\log n}\right) = \mathcal{O}(n^{-\nu})$$

for arbitrary $\nu \ge 0$.

PROOF. We first show (i). By Lemma 1 in [47] we have

$$\begin{aligned} P\left(\left\|n^{-1/2}\sum_{k=1}^{n}\mathbf{U}_k\right\|_\infty \ge \sqrt{q\log n}\right) &\le \sum_{h=1}^{d_n} P\left(\left|n^{-1/2}\sum_{k=1}^{n}U_k^{(h)}\right| \ge \sqrt{q\log n}\right) \\ &= \mathcal{O}(d_n n^{-\nu}) \end{aligned}$$

for arbitrary $\nu \ge 0$, hence the claim. Part (ii) can be shown in the same way, using Theorem 3 in [47] instead of Lemma 1. □

LEMMA 6.5. *If Assumption 2.7 is valid, then there exists a sequence of random vectors $\mathbf{U}_k^* = (U_k^{(1,*)}, \dots, U_k^{(d_n,*)})^T$ with $\mathbb{E}(\mathbf{U}_k^*) = 0$ and the same covariance structure as $\mathbf{U}_k$, such that $\mathbf{U}_k^*$ is a d_n-dependent sequence, $\max_{1\le k\le n}|U_k^{(h,*)}| = \mathcal{O}(b_n^2)$, $1 \le h \le d_n$, and*

$$P\left(n^{-1/2}\left\|\sum_{k=1}^{n}\mathbf{U}_k - \sum_{k=1}^{n}\mathbf{U}_k^*\right\|_\infty \ge v_n\right) = \mathcal{O}(n^{-\nu}),$$

where $v_n = \sqrt{n}(\log n)^{-\chi_3}$ for arbitrary ν, $\chi_3 \ge 0$.

PROOF. Put

$$\varepsilon_{k,b_n} = \varepsilon_k \mathbf{1}_{|\varepsilon_k| \le b_n} - \mathbb{E}(\varepsilon_k \mathbf{1}_{|\varepsilon_k| \le b_n}) \tag{6.2}$$

and let

$$\begin{aligned} U_{k,b_n}^{(h)} &= U_k^{(h)} \mathbf{1}_{\max_{|l| \le n} |\varepsilon_l| \le b_n} - \mathbb{E}(U_k^{(h)} \mathbf{1}_{\max_{|l| \le n} |\varepsilon_l| \le b_n}) \\ &= \varepsilon_{k,b_n} \Biggl(\sum_{j=1}^{d_n} \gamma_{h,j}^{(*)} \sum_{i=0}^{\infty} \alpha_i \varepsilon_{k-j-i,b_n} \Biggr). \end{aligned}$$

Denote with $\mathbf{U}_{k,b_n} = (U_{k,b_n}^{(1)}, \dots, U_{k,b_n}^{(d_n)})^T$; then

$$\begin{aligned} &P\Biggl(\Biggl\| \sum_{k=1}^{n} \mathbf{U}_k - \sum_{k=1}^{n} \mathbf{U}_{k,b_n} \Biggr\|_\infty \ge v_n \Biggr) \\ &\qquad \le P\Bigl(\max_{|l| \le n} |\varepsilon_l| > b_n \Bigr) + P\bigl(|\sqrt{n} \mathbb{E}(\mathbf{U}_{k,b_n}^{(h)})| \ge (\log n)^{-\chi_3} \bigr). \end{aligned}$$

Since $\mathbb{E}(\mathbf{U}_k^{(h)}) = 0$, an application of the Cauchy–Schwarz inequality yields

$$|\sqrt{n} \mathbb{E}(\mathbf{U}_{k,b_n}^{(h)})| \le \sqrt{n} \|\mathbf{U}_{k,b_n}^{(h)}\|_2 \|\mathbf{1}_{\max_{|l| \le n} |\varepsilon_l| > b_n}\|_2 = C \sqrt{n P\Bigl(\max_{|l| \le n} |\varepsilon_l| > b_n \Bigr)},$$

which by Assumption 2.7 is of the magnitude $\mathcal{O}(n^{-\nu})$, for arbitrary $\nu \ge 0$. Hence we conclude

$$P\Biggl(\Biggl\| \sum_{k=1}^{n} \mathbf{U}_k - \sum_{k=1}^{n} \mathbf{U}_{k,b_n} \Biggr\|_\infty \ge v_n \Biggr) = \mathcal{O}(n^{-\nu}). \tag{6.3}$$

Put $\mathbf{U}_{k,b_n}^{(d_n)} = (U_{k,b_n}^{(1,d_n)}, \dots, U_{k,b_n}^{(d_n,d_n)})^T$. Then

$$\mathbf{U}_{k,b_n}^{(d_n)} = \varepsilon_{k,b_n} \Biggl(\sum_{j=1}^{d_n} \gamma_{h,j}^{(*)} \sum_{i=0}^{d_n} \alpha_i \varepsilon_{k-j-i,b_n} \Biggr).$$

By Lemma 6.4 (remains valid) we have that

$$\begin{aligned} &P\Biggl(\Biggl\| \sum_{k=1}^{n} \mathbf{U}_{k,b_n} - \mathbf{U}_{k,b_n}^{(d_n)} \Biggr\|_\infty \ge v_n \Biggr) \\ &\quad \le \sum_{h=0}^{d_n} P\Biggl(\Psi(d_n)^{-1/2} \Biggl| n^{-1/2} \sum_{k=1}^{n} U_{k,b_n}^{(h)} - U_{k,b_n}^{(h,d_n)} \Biggr| \ge \Psi(d_n)^{-1/2} (\log n)^{-\chi_3} \Biggr) \\ &\quad = \mathcal{O}(n^{-\nu}) \end{aligned}$$

for arbitrary $\nu \geq 0$. Let $\{\varepsilon_k^{(h,*)}\}_{k\in\mathbb{Z}}$, $1 \leq h \leq \mathfrak{d}$, be an array of mutually independent random variables, where $\varepsilon_k^{(h,*)}$ is an independent copy of ε_{k,b_n} for each h. Then we can define the random vectors

$$U_k^{(h,*)} = U_{k,b_n}^{(h,d_n,*)} = \varepsilon_{k,b_n}\left(\sum_{j=1}^{d_n}\gamma_{h,j}^{(*)}\left[\sum_{i=0}^{d_n}\alpha_i\varepsilon_{k-j-i,b_n} + \sum_{i=d_n+1}^{\infty}\alpha_i\varepsilon_{k-j-i}^{(h,*)}\right]\right).$$

Note that due to the structure of $U_{k,b_n}^{(h,d_n,*)}$ it is clear that one may repeat all the previous arguments to derive the bound

$$n^{-1/2}\left\|\sum_{k=1}^{n}\mathbf{U}_k - \sum_{k=1}^{n}U_{k,b_n}^{(h,d_n,*)}\right\|_{\infty} = \mathcal{O}_P(n^{-\nu}). \tag{6.4}$$

Let $\sigma_n^* = \operatorname{Var}(\varepsilon_{k,b_n})$. Since $\sigma_n^* > 0$ for large enough n, the Cauchy–Schwarz inequality and Assumption 2.1 imply

$$\begin{aligned}\left|\sqrt{\sigma_n^*} - \sqrt{\sigma^2}\right|_1 &\leq C|\sigma_n^* - \sigma^2|_1 = C\left\|\varepsilon_k^2\mathbf{1}_{|\varepsilon_k|\geq b_n}\right\|_1\\ &\leq C\|\varepsilon_k^2\|_2\sqrt{P(|\varepsilon_k| \geq b_n)} = \mathcal{O}(n^{-\nu}).\end{aligned}$$

Then we obtain from the above and Lemma 6.4 (remains valid)

$$n^{-1/2}\left\|(1-\sigma^2/\sigma_n^*)\sum_{k=1}^{n}U_{k,b_n}^{(h,d_n,*)}\right\|_{\infty} = \mathcal{O}_P(n^{-\nu}). \tag{6.5}$$

Put $\mathbf{U}_k^* = (U_k^{(1,*)}, \ldots, U_k^{(d_n,*)})^T$. Then it is clear that $\max_{1\leq k\leq n}|U_k^{(h,*)}|_d = \mathcal{O}(b_n^2)$, $1 \leq h \leq d_n$, and piecing everything together, the claim follows. □

We will now construct an approximation for the random vector $\mathbf{U}_k^*$. To this end, we first divide the set of integers $\{1, 2, \ldots\}$ into consecutive blocks H_1, J_1, H_2, $J_2, \ldots.$ The blocks are defined by recursion. Fix $\delta^* > \delta > 0$, and put $m_n = \lfloor n^{\delta^*}\rfloor$. If the largest element of J_{i-1} is k_{i-1}, then $H_i = \{k_{i-1}+1, \ldots, k_{i-1}+m_n\}$ and $J_i = \{k_{i-1}+m_n+1, \ldots, d_n\}$. Let $|\cdot|$ denote the cardinality of a set. It follows from the definition of H_i, J_i that $|H_i| = m_n$ and $|J_i| = d_n$. Note that the total number of blocks is approximately $n/m_m = n^{1-\delta^*}$. Let $\mathcal{I} \subset \{0, 1, \ldots, d_n\}$ be a subset with $|\mathcal{I}| = \mathfrak{d}$, with $\mathfrak{d} = \mathcal{O}(n^{\lambda})$, $\lambda > 0$, and denote with $\sigma^2\mathbf{\Gamma}_{\mathcal{I}}$ the sub-covariance matrix of $\mathbf{U}_k^*$ restricted to the subset $\mathcal{I}$.

LEMMA 6.6. *If Assumption 2.7 is valid and* $5\lambda + 2\delta^* < 1$, *then on a possible larger probability space there exists a* $\mathfrak{d}$*-dimensional Gaussian random vector* $\mathbf{Z}$ *with covariance matrix* $n\mathbf{\Gamma}_{Z,\mathcal{I}}$, *such that*

$$P\left(\max_{h\in\mathcal{I}}\left|\mathbf{Z} - \sum_{k=1}^{n}\mathbf{U}_k^*\right| \geq v_n\right) = \mathcal{O}(\exp(-n^{\varepsilon})), \qquad \varepsilon > 0,$$

where $v_n = \sqrt{n}(\log n)^{-\chi_3}$, *for arbitrary* $\chi_3 \geq 0$, *and* $\max\|\mathbf{\Gamma}_{Z,\mathcal{I}} - \sigma^2\mathbf{\Gamma}_{\mathcal{I}}\| = \mathcal{O}(m_n^{-1})$.

PROOF. For $h \in \mathcal{I}$, let

$$\xi_k^{(h)} = \sum_{i \in H_k} U_i^{(h,*)} \quad \text{and} \quad \eta_k^{(h)} = \sum_{i \in J_k} U_i^{(h,*)}$$

and define the vectors

$$\boldsymbol{\xi}_k = (\dots, \xi_k^{(h)}, \dots)^T, \qquad h \in \mathcal{I}, \quad \text{and} \quad \boldsymbol{\eta}_k = (\dots, \eta_k^{(h)}, \dots)^T, \qquad h \in \mathcal{I}.$$

Note that per construction, we have that $\{\boldsymbol{\xi}_k\}_{k\in\mathbb{N}}$ is a sequence of independent random vectors with $|\boldsymbol{\xi}_k|_{\mathfrak{d}} = \mathcal{O}(\sqrt{\mathfrak{d}} m_n b_n^2)$. By Lemma 6.3, we can define a sequence of independent normal random vectors $\boldsymbol{\xi}_k^* = (\dots, \xi_k^{(h,*)}, \dots)^T$, $h \in \mathcal{I}$, such that for $x > 0$

$$\begin{aligned} P\Bigg(\max_{1\leq h\leq \mathfrak{d}}\Bigg|\sum_{j=1}^{n/m_n}\big(\xi_j^{(h)} - \xi_j^{(h,*)}\big)\Bigg| \geq x\Bigg) &\leq \sum_{h=1}^{\mathfrak{d}} P\Bigg(\Bigg|\sum_{j=1}^{n/m_n}\big(\xi_j^{(h)} - \xi_j^{(h,*)}\big)\Bigg| \geq x\Bigg) \\ &\leq \sum_{h=1}^{\mathfrak{d}} P\Bigg(\Bigg|\sum_{j=1}^{n/m_n}(\boldsymbol{\xi}_j - \boldsymbol{\xi}_j^*)\Bigg|_{\mathfrak{d}} \geq x\Bigg) \\ &\leq C\mathfrak{d}^2 \exp\bigg(-\frac{x}{\mathfrak{d}^{5/2} m_n b_n^2}\bigg). \end{aligned}$$

We thus obtain

$$P\Bigg(\max_{1\leq h\leq \mathfrak{d}}\Bigg|\sum_{j=1}^{n/m_n}\big(\xi_j^{(h)} - \xi_j^{(h,*)}\big)\Bigg| \geq v_n\Bigg) = \mathcal{O}(\exp(-n^{\varepsilon})), \tag{6.6}$$

and similar arguments show that there exists a sequence of independent normal random vectors $\boldsymbol{\eta}_k^* = (\dots, \eta_k^{(h,*)}, \dots)^T$, such that

$$P\Bigg(\max_{1\leq h\leq \mathfrak{d}}\Bigg|\sum_{j=1}^{n/m_n}\big(\eta_j^{(h)} - \eta_j^{(h,*)}\big)\Bigg| \geq v_n\Bigg) = \mathcal{O}(\exp(-n^{\varepsilon})).$$

Lemma 6.5 yields that $\operatorname{Var}(\eta_j^{(h,*)}) = \mathcal{O}(d_n)$ for all $j \leq m_n$, $1 \leq h \leq \mathfrak{d}$. Hence by known properties of the tails of a normal c.d.f., we obtain that

$$\begin{aligned} P\Bigg(\max_{1\leq h\leq \mathfrak{d}}\Bigg|\sum_{j=1}^{n/m_n}\eta_j^{(h),*}\Bigg| \geq v_n\Bigg) &\leq \sum_{h=1}^{\mathfrak{d}} P\Bigg(\Bigg|\sum_{j=1}^{n/m_n}\eta_j^{(h),*}\Bigg| \geq v_n\Bigg) \\ &\leq \mathfrak{d} P\big(|Z| \geq C\sqrt{d_n/m_n}(\log n)^{-\chi_3}\big) \qquad (6.7) \\ &= \mathcal{O}(\exp(-n^{\varepsilon})) \end{aligned}$$

for some $\varepsilon > 0$. This yields

$$(6.8)\qquad P\left(\max_{1\le h\le\mathfrak{d}}\left|\sum_{j=1}^{n/m_n}(\eta_j^{(h)}+\xi_j^{(h)}-\xi_j^{(h,*)})\right|\ge v_n\right)=\mathcal{O}(\exp(-n^{\varepsilon})).$$

Let $\boldsymbol{\eta}_k^{**}=(\dots,\eta_k^{(h,**)},\dots)^T$ $h\in\mathcal{I}$ be a copy of $\boldsymbol{\eta}_k^{*}$ such that $\boldsymbol{\eta}_i^{**}$ and $\boldsymbol{\xi}_j^{*}$ are independent for $i\neq j$. By the very construction of $\boldsymbol{\xi}_k,\boldsymbol{\eta}_k$, it is not hard to show that

$$\begin{aligned}\max_{i,j\in\mathcal{I}}\Bigg|&\operatorname{Cov}\left(\sum_{k=1}^{n/m_n}\eta_k^{(i)}+\xi_k^{(i)},\sum_{k=1}^{n/m_n}\eta_k^{(j)}+\xi_k^{(j)}\right)\\&-\operatorname{Cov}\left(\sum_{k=1}^{n/m_n}\xi_k^{(i,*)}+\eta_k^{(i,**)},\sum_{k=1}^{n/m_n}\xi_k^{(j,*)}+\eta_k^{(j,**)}\right)\Bigg|\\&=\mathcal{O}(n/m_n),\end{aligned}$$

which clearly implies $\max\|\boldsymbol{\Gamma}_{Z,\mathcal{I}}-\sigma^2\boldsymbol{\Gamma}_{\mathcal{I}}\|=\mathcal{O}(m_n^{-1})$. Hence, by enlarging the probability space if necessary and arguing similarly as in (6.7), we have that

$$P\left(\max_{1\le h\le\mathfrak{d}}\left|\sum_{j=1}^{n/m_n}(\xi_j^{(h)}+\eta_j^{(h)}-\xi_j^{(h,*)}-\eta_j^{(h,**)})\right|\ge v_n\right)=\mathcal{O}(\exp(-n^{\varepsilon})).$$

Finally, we obtain from the above

$$P\left(\max_{h\in\mathcal{I}}\left|\sum_{k=1}^{n}\mathbf{U}_k^{*}-\sum_{j=1}^{n/m_n}(\boldsymbol{\xi}_j^{*}-\boldsymbol{\eta}_j^{**})\right|\ge v_n\right)=\mathcal{O}(\exp(-n^{\varepsilon})),$$

which completes the proof. □

PROOF OF THEOREM 6.1. By Lemma 6.5 it suffices to establish the claim for $\{\mathbf{U}_k^{*}\}_{1\le k\le n}$. This, however, is provided by Lemma 6.6. □

Acknowledgments. The author thanks the Associate Editor and the anonymous referee for their valuable remarks and comments that considerably helped to improve the results of this paper.

REFERENCES

[1] AKAIKE, H. (1969). Fitting autoregressive models for prediction. *Ann. Inst. Statist. Math.* **21** 243–247. MR0246476

[2] AKAIKE, H. (1977). On entropy maximization principle. In *Applications of Statistics* (*Proc. Sympos.*, *Wright State Univ.*, *Dayton*, *Ohio*, 1976) 27–41. North-Holland, Amsterdam. MR0501456

[3] AN, H. Z., CHEN, Z. G. and HANNAN, E. J. (1982). Autocorrelation, autoregression and autoregressive approximation. *Ann. Statist.* **10** 926–936. MR0663443

[4] ANDERSON, T. W. (1971). *The Statistical Analysis of Time Series*. Wiley, New York. MR0283939

[5] BALKEMA, A. A. and DE HAAN, L. (1990). A convergence rate in extreme-value theory. *J. Appl. Probab.* **27** 577–585. MR1067023

[6] BARRON, A., BIRGÉ, L. and MASSART, P. (1999). Risk bounds for model selection via penalization. *Probab. Theory Related Fields* **113** 301–413. MR1679028

[7] BERK, K. N. (1974). Consistent autoregressive spectral estimates. *Ann. Statist.* **2** 489–502. Collection of articles dedicated to Jerzy Neyman on his 80th birthday. MR0421010

[8] BERKES, I., GOMBAY, E. and HORVÁTH, L. (2009). Testing for changes in the covariance structure of linear processes. *J. Statist. Plann. Inference* **139** 2044–2063. MR2497559

[9] BERMAN, S. M. (1964). Limit theorems for the maximum term in stationary sequences. *Ann. Math. Statist.* **35** 502–516. MR0161365

[10] BERTHET, P. and MASON, D. M. (2006). Revisiting two strong approximation results of Dudley and Philipp. In *High Dimensional Probability. Institute of Mathematical Statistics Lecture Notes—Monograph Series* **51** 155–172. IMS, Beachwood, OH. MR2387767

[11] BHANSALI, R. J. (1991). Consistent recursive estimation of the order of an autoregressive moving average process. *International Statistical Review/Revue Internationale de Statistique* **59** 81–96.

[12] BICKEL, P. J. and GEL, Y. R. (2011). Banded regularization of autocovariance matrices in application to parameter estimation and forecasting of time series. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **73** 711–728.

[13] BOLLERSLEV, T. (1986). Generalized autoregressive conditional heteroskedasticity. *J. Econometrics* **31** 307–327. MR0853051

[14] BOLLERSLEV, T., ENGLE, R. F. and NELSON, D. B. (1994). Arch models. In *Handbook of Econometrics, Vol. IV. Handbooks in Economics* **2** 2959–3038. North-Holland, Amsterdam. MR1315984

[15] BOX, G. E. P., JENKINS, G. M. and REINSEL, G. C. (2008). *Time Series Analysis: Forecasting and Control*, 4th ed. Wiley, Hoboken, NJ. MR2419724

[16] BROCKWELL, P. J., DAHLHAUS, R. and TRINDADE, A. A. (2005). Modified Burg algorithms for multivariate subset autoregression. *Statist. Sinica* **15** 197–213. MR2125728

[17] BROCKWELL, P. J. and DAVIS, R. A. (1991). *Time Series: Theory and Methods*, 2nd ed. Springer, New York. MR1093459

[18] DEO, C. M. (1972). Some limit theorems for maxima of absolute values of Gaussian sequences. *Sankhyā Ser. A* **34** 289–292. MR0334319

[19] ENGLE, R. (2002). Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroskedasticity models. *J. Bus. Econom. Statist.* **20** 339–350. MR1939905

[20] FOSTER, D. P. and GEORGE, E. I. (1994). The risk inflation criterion for multiple regression. *Ann. Statist.* **22** 1947–1975. MR1329177

[21] GALBRAITH, R. F. and GALBRAITH, J. I. (1974). On the inverses of some patterned matrices arising in the theory of stationary time series. *J. Appl. Probab.* **11** 63–71. MR0365959

[22] GOURIÉROUX, C. (1997). *ARCH Models and Financial Applications*. Springer, New York. MR1439744

[23] HANNAN, E. J. (1970). *Multiple Time Series*. Wiley, Sydney. MR0279952

[24] HANNAN, E. J. (1980). The estimation of the order of an ARMA process. *Ann. Statist.* **8** 1071–1081. MR0585705

[25] HANNAN, E. J. and QUINN, B. G. (1979). The determination of the order of an autoregression. *J. Roy. Statist. Soc. Ser. B* **41** 190–195. MR0547244

[26] ING, C.-K. and WEI, C.-Z. (2003). On same-realization prediction in an infinite-order autoregressive process. *J. Multivariate Anal.* **85** 130–155. MR1978181

[27] ING, C.-K. and WEI, C.-Z. (2005). Order selection for same-realization predictions in autoregressive processes. *Ann. Statist.* **33** 2423–2474. MR2211091

[28] LEEB, H. and PÖTSCHER, B. M. (2008). Model selection. In *Handbook of Financial Time Series*. Springer, New York.

[29] LÜTKEPOHL, H. (2005). *New Introduction to Multiple Time Series Analysis*. Springer, Berlin. MR2172368

[30] MALLOWS, C. L. (1964). Choosing variables in a linear regression: A graphical aid. Presented at the Central Regional Meeting of the Institute of Mathematical Statistics, Manhattan, KS 5.

[31] MALLOWS, C. L. (1966). Choosing a subset regression. Presented at the Joint Statistical Meeting, Los Angeles, CA.

[32] MCCLAVE, J. (1975). Subset autoregression. *Technometrics* **17** 213–220. MR0368359

[33] MCLEOD, A. I. and ZHANG, Y. (2008). Improved subset autoregression: With R package. *Journal of Statistical Software* **28** 1–28.

[34] NAKATSUKA, T. (1978). Regions of autocorrelation coefficients in AR(p) and EX(p) processes. *Ann. Inst. Statist. Math.* **30** 315–319. MR0514499

[35] OMEY, E. (1989). On the rate of convergence in extreme value theory. In *Stability Problems for Stochastic Models* (*Sukhumi*, 1987). *Lecture Notes in Math.* **1412** 270–279. Springer, Berlin. MR1041359

[36] PARZEN, E. (1974). Some recent advances in time series modeling. *IEEE Trans. Automat. Control* **AC-19** 723–730. System identification and time-series analysis. MR0421014

[37] QUENOUILLE, M. H. (1947). A large-sample test for the goodness of fit of autoregressive schemes. *J. Roy. Statist. Soc.* (*N.S.*) **110** 123–129. MR0025136

[38] RISSANEN, J. (1978). Modeling by shortest data description. *Automatica* **14** 465–471.

[39] SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464. MR0468014

[40] SHAO, J. (1997). An asymptotic theory for linear model selection. *Statist. Sinica* **7** 221–264. With comments and a rejoinder by the author. MR1466682

[41] SHIBATA, R. (1976). Selection of the order of an autoregressive model by Akaike's information criterion. *Biometrika* **63** 117–126. MR0403130

[42] TONG, H. (1977). Some comments on the canadian lynx data. *J. Roy. Statist. Soc. Ser. A* **140** 432–436.

[43] WALKER, A. M. (1952). Some properties of the asymptotic power functions of goodness-of-fit tests for linear autoregressive schemes. *J. Roy. Statist. Soc. Ser. B.* **14** 117–134. MR0050239

[44] WALKER, G. (1931). On periodicity in series of related terms. *Monthly Weather Review* **59** 277–278.

[45] WHITTLE, P. (1952). Tests of fit in time series. *Biometrika* **39** 309–318. MR0052743

[46] WHITTLE, P. (1963). On the fitting of multivariate autoregressions, and the approximate canonical factorization of a spectral density matrix. *Biometrika* **50** 129–134. MR0161430

[47] WU, W. B. (2009). An asymptotic theory for sample covariances of Bernoulli shifts. *Stochastic Process. Appl.* **119** 453–467. MR2493999

[48] WU, W. B. and XIAO, H. (2011). Asymptotic inference of autocovariances of stationary processes. Available at arXiv:1105.3423.

[49] YULE, U. G. (1927). On a method of investigating periodicities in disturbed series, with special reference to wolfer's sunspot numbers. *Phil. Trans. R. Soc. Lond. A* **226** 267–298.

INSTITUTE OF STATISTICS
GRAZ UNIVERSITY OF TECHNOLOGY
MÜNZGRABENSTRASSE 11, A-8010 GRAZ
AUSTRIA
E-MAIL: m0ritz@yahoo.com