Non–Asymptotic Bounds for Autoregressive Time Series Modeling

Alexander Goldenshluger* Assaf Zeevi†

Published 2001, *Annals of Statistics*

Abstract

The subject of this paper is autoregressive (AR) modeling of a stationary, Gaussian discrete time process, based on a finite sequence of observations. The process is assumed to admit an AR(∞) representation with exponentially decaying coefficients. We adopt the nonparametric minimax framework and study how well can the process be approximated by a finite order AR model. A lower bound on the accuracy of AR approximations is derived, and a non-asymptotic upper bound on the accuracy of the regularized least squares estimator is established. It is shown that with a "proper" choice of the model order, this estimator is minimax optimal in order. These considerations lead also to a non-asymptotic upper bound on the mean squared error of the associated one step predictor. A numerical study compares the common model selection procedures to the minimax optimal order choice.

1 Introduction

The standard methods for estimating parameters of time series are based on the assumption that the observations come from an autoregressive (AR), moving average (MA), or mixed (ARMA) model of known orders. This assumption can be rarely justified in practice, and the less stringent assumption is that the time series data are observations from a linear stationary process. A common approach to modeling linear stationary processes is based on an AR approximation. In this framework a finite order AR model is fitted to the observations. The order of the AR model should provide an "optimal" finite AR approximation to the process, and it is usually chosen by selection procedures based on the data. This nonparametric AR approach to modeling linear stationary processes has been investigated by Shibata (1980), Bhansali (1981, 1986), An *et al.* (1982), and Hannan and Kavalieris (1986).

Shibata (1980) considered the problem of predicting a Gaussian infinite–order AR process by fitting a finite AR model. The notion of optimality for the model selection procedure

*Alexander Goldenshluger is with Department of Statistics, The University of Haifa, Haifa 31905, Israel
†Assaf Zeevi is with the Information Systems Lab, Stanford University, Stanford, CA. 94305-9510

proposed by Shibata (1980) is based on an asymptotic lower bound on the mean squared prediction error. Specifically, the procedure is asymptotically efficient if it attains the lower bound asymptotically. Shibata (1980) also established that the final prediction error (FPE) [Akaike (1970)] and the AIC [Akaike (1974)] criteria are asymptotically efficient in the above sense, provided that the linear process does not degenerate to a finite order autoregression. A similar result has been obtained by Bhansali (1986) for the AR transfer function criterion (CAT) proposed by Parzen (1974).

Another motivation for fitting an AR model is the estimation of the spectral density function. Berk (1974) used AR approximation to estimate the spectral density of a linear process. It was shown there that the order of the approximating AR model should increase with the number of observations to ensure the consistency of the associated spectral density estimator. Shibata (1981) suggested another definition of selection procedures optimality which is based on an asymptotic lower bound for the relative integrated squared error in estimating the spectral density function. It was shown there that the FPE and AIC criteria are asymptotically efficient in this sense, provided that the linear process does not degenerate to a finite order autoregression. Similar results for the CAT criterion have been obtained by Bhansali (1986). Some recent results on AR approximation can be found in Gerencsér (1992) and Bülmann (1995).

In spite of the fact that the FPE, AIC, and CAT criteria are asymptotically efficient as described above, the finite sample behavior of these selection procedures is not so clear. The definitions of optimality adopted in Shibata (1980, 1981) and Bhansali (1986) are essentially asymptotic. The assumption that the underlying linear process does not degenerate to a finite autoregression is also based on asymptotic considerations. If this assumption is violated, the AIC and FPE overestimate the true model order, and a different penalty term is called for. In particular, by penalizing each parameter by a factor of $\ln n$, with n being the sample size, one obtains the minimum description length (MDL) principle of Rissanen (1983), and the BIC criterion of Schwarz (1978). These criteria lead to consistent estimation of the model dimension in the case of an underlying finite order autoregression. However, if the underlying process does not degenerate to a finite autoregression, they are not asymptotically efficient in the aforementioned sense [cf. the discussion in Shibata (1980, pp. 161)]. Moreover, even if the underlying "true" linear process does not degenerate to a finite order autoregression, the coefficients in its AR(∞) representation can be small. In these situations, effectively the model is "close" to being finite dimensional, and the behavior of the asymptotic efficient procedures can be poor even for "large" sample sizes. Several interesting questions arise in this context. Given a fixed number of observations from a linear process, how well can the underlying process be modeled using a finite order autoregression? How can the finite sample behavior of selection procedures be assessed? It is evident that another notion of optimality is needed in order to address these questions.

In this paper we propose to use the nonparametric minimax approach to measuring the accuracy of an AR approximation. This framework is very common in nonparametric estimation problems such as nonparametric regression, density estimation and spectral density estimation. According to this methodology, we assume that the linear process belongs to a certain class, and the quality of an approximating AR model (and the associated one step predictor) is measured by its worst-case modeling (respectively, prediction) risk over the class. Establishing non–asymptotic upper and lower bounds on the risk, one can assess accuracy of an estimator. Throughout the paper we consider the class of linear processes admitting an AR(∞) representation with exponentially decaying coefficients. The practical importance of the class follows from the fact that it includes (but is not limited to) all causal invertible ARMA(p, q) processes. We derive a non–asymptotic lower bound on the accuracy of an AR approximation, and show that the least squares estimator with a "proper" choice of the order is minimax optimal in order. These considerations lead also to a non–asymptotic upper bound on the mean squared error of the associated one step predictor. Further, we present some numerical examples comparing common selection procedures (FPE, AIC and MDL) to the minimax optimal one. We note that our derivation is based on an exponential inequality on deviations of the sample covariances from their expectations; these results are of independent interest. The same technique has been used in Goldenshluger (1998) for derivation of non–asymptotic bounds in estimating impulse response sequences of linear dynamic systems.

The rest of the paper is organized as follows. In Section 2 we state formally the problem of nonparametric AR approximation in the minimax framework. Section 3 describes the construction of the estimator, and presents main results. In Section 4 we present our numerical examples. Some remarks are collected in Section 5. The proofs are given in Apendices A, B, and C.

2 Minimax framework and overview of results

Let $(X_t)_{t\in\mathbf{Z}}$ be a real–valued, purely nondeterministic, Gaussian stationary process with zero mean, $E|X_t|^2 = 1$, spectral density function $f(\lambda)$, $\lambda \in [-\pi, \pi]$, and covariance function $\gamma(k)$, $k \in \mathbf{Z}$. According to the Wold decomposition theorem, $(X_t)_{t\in\mathbf{Z}}$ can be represented as an MA(∞) process

$$X_t = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j}, \quad \psi_0 = 1, \quad \sum_{j=0}^{\infty} \psi_j^2 < \infty, \tag{1}$$

where $\{\varepsilon_t\}_{t\in\mathbf{Z}}$ is a sequence of independent Gaussian innovations with $E\varepsilon_t = 0$ and $E\varepsilon_t^2 = \sigma_\varepsilon^2$. Assume that the MA transfer function $\Psi(z) = \sum_{j=0}^\infty \psi_j z^j$ has no zeros in the unit disc $|z| \le 1,\ z \in \mathbf{C}$, and $\sum_{j=0}^\infty |\psi_j| < \infty$; then the linear process $(X_t)_{t\in\mathbf{Z}}$ can also be represented as an invertible AR(∞) process

$$X_t = \sum_{j=1}^{\infty} \phi_j X_{t-j} + \varepsilon_t, \quad t \in \mathbf{Z}, \tag{2}$$

where the coefficients $\phi_j,\ j = 1, \ldots, \infty$ are given by $1/\Psi(z) = 1 - \sum_{j=1}^\infty \phi_j z^j$. Given observations $X_1, \ldots, X_n$ from the process $(X_t)_{t\in\mathbf{Z}}$, we are interested in modeling $(X_t)_{t\in\mathbf{Z}}$ and predicting the future value X_{n+1}. The representation (2) motivates the use of AR approximation to approach the problems of modeling and prediction.

Assume that the process $(X_t)_{t\in\mathbf{Z}}$ belongs to a certain family, and the quality of an approximating AR model (and the associated one step predictor) is measured by the worst-case modeling (respectively, prediction) error over the family. The problem of modeling the process $(X_t)_{t\in\mathbf{Z}}$ by a finite order AR model is identical to estimating the corresponding coefficient sequence $\phi = (\phi_1, \phi_2, \ldots)$ in the AR(∞) representation of $(X_t)_{t\in\mathbf{Z}}$. More formally, let $\mathcal{H}$ be a family of stationary Gaussian processes $(X_t)_{t\in\mathbf{Z}}$ with zero mean and unit variance, admitting an AR(∞) representation (2). Let $\hat\phi = \hat\phi(X_1, \ldots, X_n)$ be an estimator of the sequence $\phi = (\phi_1, \phi_2, \ldots)$; then quality of the estimator $\hat\phi$ is measured by its maximal risk over $\mathcal{H}$

$$\mathcal{R}_m[\hat\phi, \mathcal{H}] := \sup_{(X_t)\in\mathcal{H}} \Big[E\|\hat\phi - \phi\|^2\Big]^{1/2},$$

where $\|\cdot\|$ is the standard ℓ_2 norm in the space of sequences. The *minimax* estimator $\hat\phi_* = \hat\phi_*(X_1, \ldots, X_n)$ is the one minimizing the maximal risk

$$\mathcal{R}_m^*[n, \mathcal{H}] := \inf_{\hat\phi} \mathcal{R}_m[\hat\phi, \mathcal{H}] = \inf_{\hat\phi} \sup_{(X_t)\in\mathcal{H}} \Big[E\|\hat\phi - \phi\|^2\Big]^{1/2},$$

where the infimum is taken here over all possible estimators. Typically, the minimax estimators cannot be constructed, therefore, as usual in nonparametric estimation, we will be interested in *optimal in order* estimators for which

$$\mathcal{R}_m[\hat\phi, \mathcal{H}] \le C(n)\mathcal{R}_m^*[n, \mathcal{H}], \quad \sup_n C(n) < \infty. \tag{3}$$

Similarly, in the problem of prediction of X_{n+1} using observations $X_1, \ldots, X_n$ we will measure the accuracy of a prediction method $\hat X_{n+1}(X_1, \ldots, X_n)$ by its maximal prediction error over $\mathcal{H}$

$$\mathcal{R}_p[\hat X_{n+1}, \mathcal{H}] := \sup_{(X_t)\in\mathcal{H}} \Big[E(\hat X_{n+1} - X_{n+1})^2 - \sigma_\varepsilon^2\Big].$$

The minimax prediction error is defined as the infimum of the maximal prediction error, over all possible prediction methods

$$\mathcal{R}_p^*[n,\mathcal{H}] := \inf_{\hat{X}_{n+1}} \mathcal{R}_p[\hat{X}_{n+1},\mathcal{H}] = \inf_{\hat{X}_{n+1}} \sup_{(X_t)\in\mathcal{H}} \Big[E(\hat{X}_{n+1}-X_{n+1})^2 - \sigma_\varepsilon^2\Big].$$

In what follows, we will be interested in *optimal in order* predictors for which (3) holds with $\mathcal{R}_m$ replaced by $\mathcal{R}_p$.

Throughout the paper we restrict attention to the following family $\mathcal{H}_\rho(l,L)$ of stationary Gaussian processes $(X_t)_{t\in\mathbf{Z}}$ satisfying $EX_t=0$, $E|X_t|^2=1$. For given finite real numbers $\rho>1$, $0<l<1$ and $L>1$, define $\mathcal{H}_\rho(l,L)$ as

$$\mathcal{H}_\rho(l,L) := \Big\{(X_t) : 0<l\le|\Psi(z)|\le L, \text{ for } |z|\le\rho\Big\},$$

where $\Psi(\cdot)$ is the MA(∞) transfer function. In words, the MA(∞) transfer function of the process $(X_t)_{t\in\mathbf{Z}}\in\mathcal{H}_\rho(l,L)$ is analytic in an open set containing the disc $|z|\le\rho$, and bounded from above and below by constants L and l, respectively. The class $\mathcal{H}_\rho(l,L)$ contains Gaussian stationary processes with spectral density function $f(\lambda)$ bounded away from zero and infinity, which can be continued analytically over the interior of the strip $\{(x+iy)\in\mathbf{C} : |y|<\ln\rho\}$ in the complex plane. The parameters l and L in the definition of $\mathcal{H}_\rho(l,L)$ guarantee uniform lower and upper bounds on the spectral density function. This, in turn, implies uniform bounds on the eigenvalues of the covariance matrices of all orders [cf. Grenander and Szegö (1984)]. Practical importance of the class $\mathcal{H}_\rho(l,L)$ stems from the fact that it contains causal invertible ARMA(p,q) processes with proper restrictions on the magnitude of the coefficients. For example, all MA(1) processes with the coefficient $|\psi_1|\le\rho^{-1}\min\{1-l,L-1\}$ belong to $\mathcal{H}_\rho(l,L)$. The processes from $\mathcal{H}_\rho(l,L)$ admit AR(∞) representation with uniformly bounded exponentially decaying coefficients.

Remark 1 We note here a simple imbedding relationship between classes $\mathcal{H}_\rho(l,L)$ with different parameters: $\mathcal{H}_\rho(l,L)\subseteq\mathcal{H}_r(l,L)$, $\forall\rho\ge r>1$. As we shall see, the only important parameter for constructing a rate optimal AR estimator (predictor) is ρ.

Remark 2 The classes of analytic functions are quite standard in nonparametric estimation problems. Our class is similar to those in Golubev and Levit (1996) and Golubev, Levit and Tsybakov (1997). In the context of spectral density estimation, a closely related class of processes was considered by Efromovich (1998).

The main contributions of this paper are the following. We study how well processes $(X_t)_{t\in\mathbf{Z}}\in\mathcal{H}_\rho(l,L)$ can be approximated by a finite order AR model, obtaining a lower bound on the minimax risk $\mathcal{R}_m^*[n, \mathcal{H}_\rho(l, L)]$. We prove that if the sample size n is large enough then

$$\mathcal{R}_m^*[n, \mathcal{H}_\rho(l, L)] \geq K(l, L)\left(\frac{\rho - 1}{\rho}\right)\frac{1}{\sqrt{\ln \rho}}\sqrt{\frac{\ln n}{n}},$$

where the constant $K(l, L)$ depends on l and L only. A non-asymptotic upper bound on the maximal risk of the regularized least squares estimator is derived. We show that the least squares estimator associated with the order $d_* = \lfloor (2 \ln \rho)^{-1} \ln n \rfloor$ of the approximating AR model is optimal in order in the sense of inequality (3). These results have immediate implications for the prediction problem. In particular, we derive a non-asymptotic upper bound on $\mathcal{R}_p[\hat{X}_{n+1}, \mathcal{H}_\rho(l, L)]$ for the corresponding one step predictor and argue that the predictor associated with the order d_*, is essentially minimax optimal in order. The non-asymptotic bounds we obtain are based on exponential inequalities on deviations of sample covariances from their expectations; these results are of independent interest. Further, through numerical examples we compare small samples behavior of some common order selection procedures to the minimax optimal choice d_*. In particular, simulating an MA(1) process, we found that for moderate values of ρ, i.e., when the zeros are not 'too close' to the unit disc, the AIC and FPE lead to an order selection that is comparable to the minimax optimal one. However, if ρ is close to unity then the AIC and FPE tend to select a smaller model order than the minimax optimal one. The MDL turns out to be slightly more conservative than the other methods, with the differences becoming marginal for larger values of ρ.

3 Main results

Consider the following estimate of the AR sequence $\phi = (\phi_1, \phi_2, \ldots)$. Fix a natural number d, and define

$$\theta^d = (\phi_1, \ldots, \phi_d)', \quad Z_t = (X_{t-1}, \ldots, X_{t-d})'.$$

We estimate θ^d by the regularized least squares method:

$$\hat{\theta}^d = \left(\frac{1}{n}\sum_{t=1}^{n} Z_t Z_t' + n^{-1} I_d\right)^{-1}\left(\frac{1}{n}\sum_{t=1}^{n} X_t Z_t\right), \qquad \hat{\theta}^d = (\hat{\phi}_1, \ldots, \hat{\phi}_d)', \tag{4}$$

where I_d is the identity $d \times d$ matrix. The corresponding estimate $\hat{\phi}$ of the sequence $\phi = (\phi_1, \phi_2, \ldots)$ is given by

$$\hat{\phi} = (\hat{\phi}_1, \hat{\phi}_2, \ldots, \hat{\phi}_d, 0, 0, \ldots) \tag{5}$$

and the one-step predictor $\hat{X}_{n+1}^d$ based on $\hat{\phi}$ is defined by

$$\hat{X}_{n+1}^d = \sum_{j=1}^{d} \hat{\phi}_j X_{n+1-j}. \tag{6}$$

The reason why we consider a regularized version of the least squares estimate is that we are interested in a non–asymptotic upper bound on the expected value of the squared modeling (prediction) error. For this purpose we have to control the norm of the random matrix $(n^{-1}\sum_{t=1}^n Z_t Z_t')^{-1}$. Without the regularization term the matrix $n^{-1}\sum_{t=1}^n Z_t Z_t'$ can be singular with non–zero probability for every fixed n. We note also that the vectors $Z_t,\ t = 1,\dots,n$ defined above can involve X_t with $t \le 0$. In this case we suppose that the corresponding components of the vectors in (4) are replaced by zero. It should be stressed however that in our analysis we do not assume that $X_t = 0$ for $t \le 0$.

3.1 Accuracy of AR approximation

We are now ready to study the quality of an AR approximation of the stationary Gaussian process $(X_t)_{t\in\mathbf{Z}} \in \mathcal{H}_\rho(l, L)$.

Theorem 1 *Let*

$$M = 1 + \frac{L\rho}{l(\rho-1)}, \quad r = 1 + \frac{1}{\ln\rho}. \tag{7}$$

Suppose that n and d satisfy the following conditions:

$$\frac{n}{(\ln n)^5} \ge c_1 d(rM)^5, \qquad \sqrt{\frac{n}{\ln n}} \ge c_2 (L/l)^2 d^2 \sqrt{rM}, \tag{8}$$

where c_1 and c_2 are absolute constants which can be specified explicitly. Then for the estimate (4)-(5) one has

$$\mathcal{R}_m[\hat{\phi}, \mathcal{H}_\rho(l,L)] \le K_1(l,L)\left(\frac{1}{n(\rho-1)} + \frac{\sqrt{d}}{\rho^d(\rho-1)} + \sqrt{\frac{d}{n}}\right), \tag{9}$$

where $K_1(l, L)$ depends on l and L only.

Remark 3 Accuracy of the AR approximation is limited by two factors. First, we approximate the process by a finite order AR model. The resulting *approximating error* (second term in the right hand side of (9)) becomes smaller as the order d of the approximating AR model increases. Second, we estimate parameters of the approximating model. The resulting *estimating error* (third term in the right hand side of (9)) grows as the order of the approximating model increases. The order d is viewed as a "smoothing parameter" that controls a trade–off between the approximation and estimation errors. The first term in the right hand side of (9) is due to the use of a regularized version of the least squares estimator.

The following statement is an immediate consequence of Theorem 1.

Corollary 1 *Let n be large enough so that*

$$\frac{n}{(\ln n)^6} \geq c_1 r^6 M^5, \quad \frac{\sqrt{n}}{(\ln n)^{5/2}} \geq c_2 \left(\frac{L}{l \ln \rho}\right)^2 \sqrt{rM}. \tag{10}$$

Then for the least squares estimator (4)-(5) associated with the choice $d_* \triangleq \lfloor (2\ln\rho)^{-1} \ln n \rfloor$ one has

$$\mathcal{R}_m[\hat{\phi}_*, \mathcal{H}_\rho(l, L)] \leq K_2(l, L) \left(\frac{\rho}{\rho - 1}\right) \frac{1}{\sqrt{\ln \rho}} \sqrt{\frac{\ln n}{n}}, \tag{11}$$

where $K_2(l, L)$ depends on l and L only.

The next step in the analysis is to determine the limits of achievable accuracy for AR approximation. The following theorem gives a lower bound on approximation of $(X_t)_{t \in \mathbf{Z}} \in \mathcal{H}_\rho(l, L)$ by a finite order AR model.

Theorem 2 *Let n be large enough so that for some constant K_3 depending on l and L only*

$$\ln n \geq K_3(l, L) \ln \rho. \tag{12}$$

Then

$$\mathcal{R}^*_m[n, \mathcal{H}_\rho(l, L)] \geq K_4(l, L) \left(\frac{\rho - 1}{\rho}\right) \frac{1}{\sqrt{\ln \rho}} \sqrt{\frac{\ln n}{n}} \quad . \tag{13}$$

Theorem 2 and Corollary 1 imply that the least squares estimator (4)-(5) associated with the order $d_* = \lfloor (2 \ln \rho)^{-1} \ln n \rfloor$ is optimal in order in the sense of inequality (3). It is interesting to note that for the class of spectral densities corresponding to processes closely related to the class $\mathcal{H}_\rho(l, L)$, this choice of the order leads to the asymptotically minimax spectral estimate [Efromovich (1998)].

3.2 Prediction via AR approximation

In this section we establish a non–asymptotic upper bound on accuracy of the one step predictor which is based on AR approximation. To simplify analysis we assume that the estimate $\hat{\phi}$ of the sequence ϕ is based on $\lfloor n/2 \rfloor$ first observations $(X_1, \ldots, X_{\lfloor n/2 \rfloor})$ only. The assumption of this type is quite usual in investigating accuracy of prediction methods based on the estimated parameters. For instance, Shibata (1980) assumed the more stringent assumption that we have two independent realizations of the linear process: the first time series is used for estimating parameters, and then the estimated parameters are used to predict the second time series.

The associated one-step predictor is defined in (6). We note also that Theorem 1 remains unaltered for the estimate in question with n replaced by $n/2$.

Theorem 3 *Let $n > 4d$ and (8) hold with some absolute constants c_1 and c_2. Then one has*

$$\mathcal{R}_p[\hat{X}^d_{n+1}, \mathcal{H}_\rho(l,L)] \leq K_5(l,L)\left(\frac{1}{n^2(\rho-1)^2} + \frac{d}{\rho^{2d}(\rho-1)^2} + \frac{d}{n}\right)\left(1 + \frac{d\rho^{-2d}}{(\rho-1)^2}\right), \tag{14}$$

where $K_5(l,L)$ depends on l and L only.

Now, choosing the model order d we obtain the prediction bounds.

Corollary 2 *Let (10) hold with some absolute constants c_1 and c_2, and $n(\ln n)^{-1} \geq 2(\ln\rho)^{-1}$. Then for the one-step predictor $\hat{X}^d_{n+1}$ associated with the choice*

$$d_* = \lfloor (2\ln\rho)^{-1}\ln n \rfloor$$

one has

$$\mathcal{R}_p[\hat{X}^{d_*}_{n+1}, \mathcal{H}_\rho(l,L)] \leq K_6(l,L)\left(\frac{\rho}{\rho-1}\right)^2\left(\frac{1}{\ln\rho}\right)\frac{\ln n}{n}, \tag{15}$$

where $K_6(l,L)$ depends on l and L only.

Referring back to (9), we see that the upper bound on prediction accuracy given in (14) behaves as the square of accuracy of modeling. This is not surprising given the construction of the one step predictor (6); clearly the resulting accuracy is determined by the quality of modeling via the AR approximation. In fact, one can argue that the predictor $X^{d_*}_{n+1}$ is optimal in order. For the sake of simplicity assume as in Shibata (1980) that we have two independent copies $Y_1,\ldots,Y_n$ and $X_1,\ldots,X_n$ of the same linear process from $\mathcal{H}_\rho(l,L)$. Our goal is to predict X_{n+1}. Let $\hat{X}_{n+1}$ be an arbitrary prediction method for X_{n+1} based on the observations $Y_1,\ldots,Y_n$. Then $\hat{X}_{n+1}$ can be decomposed into a sum of two random variables $\hat{X}'_{n+1}$ and $\hat{X}''_{n+1}$ such that $\hat{X}'_{n+1}$ is the projection of $\hat{X}_{n+1}$ on $\overline{\mathrm{sp}}\{X_n, X_{n-1},\ldots\}$, and $\hat{X}''_{n+1}$ is orthogonal to $\overline{\mathrm{sp}}\{X_n, X_{n-1},\ldots\}$. Therefore

$$\begin{aligned}\mathcal{R}^*_p[n,\mathcal{H}_\rho(l,L)] &\geq \sup_{(X_t)\in\mathcal{H}_\rho(l,L)} E\left|\hat{X}'_{n+1} - \sum_{j=1}^\infty \phi_j X_{n+1-j}\right|^2 \\ &= \sup_{(X_t)\in\mathcal{H}_\rho(l,L)} E\left|\sum_{j=1}^\infty (\hat{\phi}_j - \phi_j)X_{n+1-j}\right|^2,\end{aligned}$$

where $\hat{\phi}_j$, $j=1,2,\ldots$ are measurable functions of $Y_1,\ldots,Y_n$. Hence

$$\mathcal{R}^*_p[n,\mathcal{H}_\rho(l,L)] \geq K_7(l,L)\sup_{\phi\in\mathcal{H}_\rho(l,L)} E\|\hat{\phi}-\phi\|^2,$$

where the supremum is taken over all sequences $\phi = (\phi_1,\phi_2,\ldots)$ that define, through the AR(∞) representation, the linear processes from $\mathcal{H}_\rho(l,L)$. Then the lower bound on the minimax prediction risk follows from Theorem 2.

4 Numerical examples

The choice of model order, $d(n) = O(\ln n)$, arises in Shibata (1980), and more recently in Hannan, Kavalieris (1986) and Gerencsér (1992). In particular, Shibata (1980) showed that the data-driven order selector based on the final prediction error (FPE) behaves asymptotically as $O(\ln n)$ for the class of processes, similar to $\mathcal{H}_\rho(m, L)$. To investigate the practical impact of the above results, we compare the common model selection strategies (AIC, FPE, and MDL) to the minimax optimal rule through a simple numerical example.

Consider the following MA(1) process

$$X_t = \varepsilon_t + \psi_1 \varepsilon_{t-1}$$

with $\{\varepsilon_t\}$ a sequence of i.i.d. standard Gaussian random variables. We focus our attention on three particular cases, namely $\psi_1 = 0.1, 0.5, 0.9$, and the corresponding 'margin of stability' $\rho = 10, 2, 1.1111$. This range of values will illustrate the sensitivity of the order selection methods to the moduli of the zeros of the transfer function $\Psi(\cdot)$. Suppose we are given n consecutive observations $X_1, X_2, \ldots, X_n$ from the process (X_t). The selection procedures are defined [following the definitions in Shibata (1980)] as

$$\begin{aligned} \mathrm{AIC}(d) &:= (n + 2d)\hat{\sigma}_d^2 \\ \mathrm{FPE}(d) &:= n((n + d)/(n - d))\hat{\sigma}_d^2 \\ \mathrm{MDL}(d) &:= (n + d\ln n)\hat{\sigma}_d^2 \end{aligned}$$

with

$$\hat{\sigma}_d^2 := \frac{1}{n-d} \sum_{t=d+1}^{n} \left(X_t - \sum_{j=1}^{d} \hat{\phi}_j X_{t-j} \right)^2 .$$

Recall also the minimax optimal order choice from Corollary 1: $d_* = \lfloor (2 \ln \rho)^{-1} \ln n \rfloor$.

The experiment was conducted by simulating 100 sample paths from the process, for each trial run a model order was selected using the three procedures, for sample sizes $n = 100, 500, 1000, 5000, 10000, 50000$ and 100000. Finally, we averaged out the selected orders over the 100 runs. The graphs in Figure 1 depict the behavior of the different order selection procedures.

A close look at Figure 1 reveals that the AIC and FPE, which are known to be asymptotically equivalent, behave in an almost identical way also for small values of sample size. The MDL leads to a choice that is more conservative than AIC and FPE, with this behavior being more pronounced for the case of small ρ. For the case of large ρ, the all three criteria are roughly the same as the minimax optimal choice. The case of moderate ρ depicts a

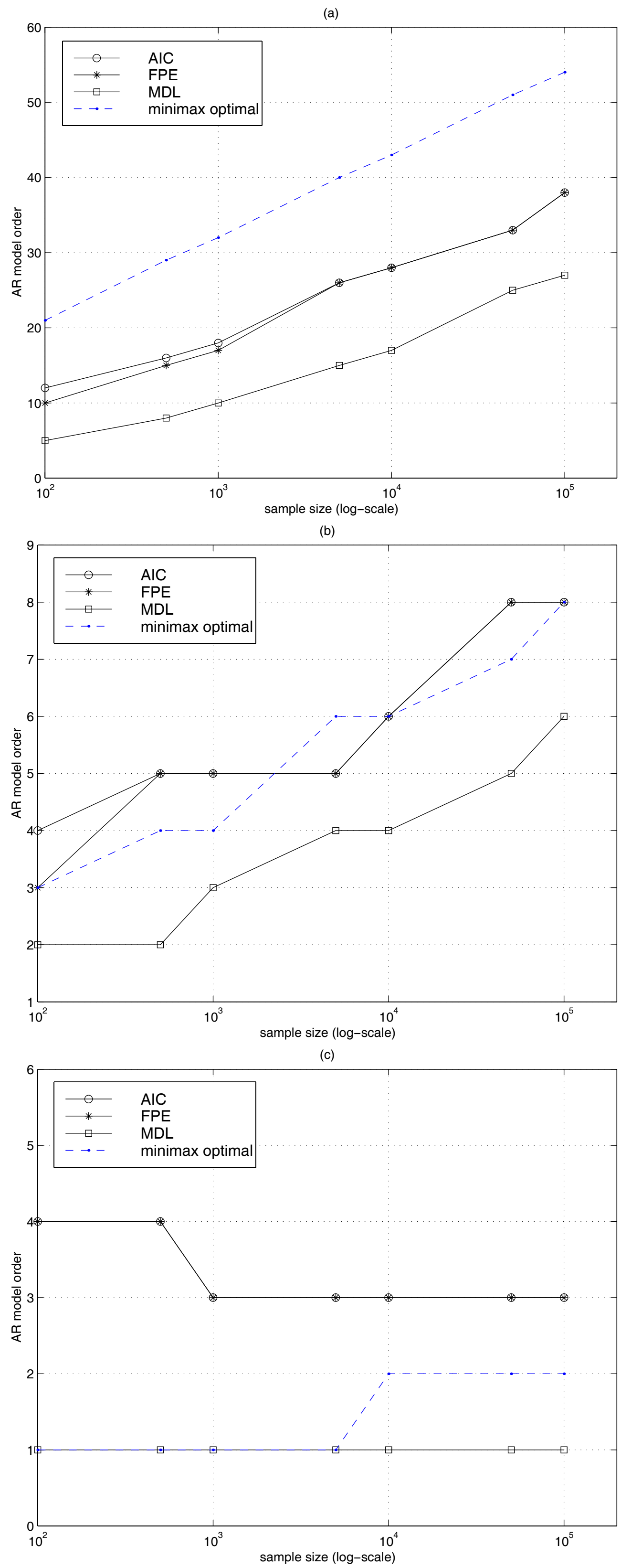

Figure 1: Model order selected by different procedures plotted against the sample size (log scale); (a) $\rho = 1.1111$, (b) $\rho = 2$, and (c) $\rho = 10$.

behavior of AIC and FPE which is quite on par with the minimax optimal choice. However, if ρ is close to unity, then the AIC and FPE tend to select a smaller model order than the minimax optimal one. It is interesting to note that all procedures lead to an order selection that exhibits logarithmic–like growth in the sample size, even for small sample sizes. This behavior is consistent with the asymptotic logarithmic growth of the order selected by AIC, and FPE [cf. Shibata (1980, Example 4.1)], and for MDL [cf. Gerencér (1992, Theorem 4)].

To summarize the results, we observe that an infinite order AR model that is closer to a parametric (finite dimensional) model gives rise to an order selection that is "close" to minimax optimal by all three methods. The case of more slowly decaying coefficients (larger ψ and ρ closer to unity, respectively) reveals that AIC and FPE 'underestimate' with MDL being even more conservative. We note in passing that similar numerical results were obtained for more complicated ARMA structures.

5 Discussion

1. The method of AR approximations is quite common for spectral density estimation in time series analysis [see, e.g., Berk (1974), Shibata (1981), Parzen (1983) among many others]. The minimax optimal model order ($d_* = \lfloor \ln n/(2\ln\rho) \rfloor$) for the AR approximation is also the optimal choice for spectral density estimation, and gives rise to the same convergence rates over the class $\mathcal{H}_\rho(l, L)$. It is worth noting, however, that spectral density estimation and AR approximation are not equivalent in the sense of comparison of experiments. Specifically, assume that the process belongs to the class of all invertible MA(q) processes whose MA–transfer function has no zeros inside the disc $|z| \le \rho$, $\rho > 1$. This class is a subset of $\mathcal{H}_\rho(l, L)$ with proper l and L. The spectral density of such a process can be estimated with the parametric rate $O(\sqrt{q/n})$, while the accuracy of the AR approximation is $O(\sqrt{\ln n/(n \ln\rho)})$. An important impilication of this fact is that even if a stationary process is approximated by an AR model with high accuracy, the corresponding spectral density estimate may be poor.

2. The nonparametric minimax approach, as applied to AR approximation, provides a useful criterion for assessment of finite sample behavior of selection methods. Within this approach, optimal selection methods are specified, and achievable lower bounds on the estimation accuracy are calculated. Note, however, that implementation of the minimax optimal rule requires a priori information on the parameter ρ of the class $\mathcal{H}_\rho(l, L)$. Developing adaptive selection rules with good minimax properties remains a challenging open problem. We conjecture that in the adaptive setting the rates of convergence for AR

approximation remain unchanged.

3. Throughout the paper we assume that the process $(X_t)_{t\in\mathbf{Z}}$ is Gaussian. This assumption is used to simplify the derivation of the exponential inequalities on the covariance estimates (Lemma 2 below). In addition, it facilitates the evaluation of higher order moments. The main results of the paper can be obtained under moment growth restrictions accompanied with some requirements ensuring exponential mixing properties of the process $(X_t)_{t\in\mathbf{Z}}$.

4. The family $\mathcal{H}_\rho(l,L)$ allows for the processes admitting AR(∞) representation with exponentially decaying coefficients. It seems that the exponential decay of the coefficients is essential for the exponential inequalities we derive. The techniques advocated in Lemma 6 and Lemma 7 below preclude polynomially decaying sequences. Thus, this restriction is a direct consequence of the limitations of our machinery. An interesting problem is to study rates of AR approximation for other classes of stationary, e.g., with polynomially decaying AR coefficients.

A Preliminary results

We collect here several preliminary results which will be used repeatedly in the subsequent proofs.

We start with establishing a relation between the properties of the sequences $\gamma(k)$, ψ_j, and ϕ_j to the class $\mathcal{H}_\rho(l,L)$. Let us define $\Gamma_d \equiv \{\gamma(i-j)\}_{i,j=1,\dots,d}$ for every natural number d.

Lemma 1 *Let $(X_t)_{t\in\mathbf{Z}} \in \mathcal{H}_\rho(l,L)$; then*

$$|\psi_j| \le L\rho^{-j}, \qquad |\phi_j| \le l^{-1}\rho^{-j}, \qquad j=1,2,\dots \tag{16}$$

In addition, we have

$$L^{-2} \le \sigma_\epsilon^2 \le l^{-2}, \tag{17}$$

$$|\gamma(k)| \le (L/l)^2 \frac{\rho^2}{\rho^2-1}\rho^{-|k|}, \quad k\in\mathbf{Z}, \tag{18}$$

and for any d

$$(l/L)^2 \le \|\Gamma_d\| \le (L/l)^2, \qquad (l/L)^2 \le \|\Gamma_d^{-1}\| \le (L/l)^2, \tag{19}$$

where $\|\cdot\|$ stands for the standard Euclidean norm of a matrix.

Proof By definition of the class $\mathcal{H}_\rho(l, L)$, $\Psi(z)$ is analytic in the open disc $|z| < \rho$, and $|\Psi(z)| \le L$. Therefore the announced bound on $|\psi_j|$ follows immediately from the Cauchy estimates for the derivatives of $\Psi(z)$ [see, e.g., Rudin (1964, pp. 229)]. Further, note that

$$L^{-1} \le |\Phi(z)| = 1/|\Psi(z)| \le l^{-1}, \ \text{ for } |z| < \rho.$$

Again applying the Cauchy estimates we obtain (16).

Note that $f(\lambda) = (2\pi)^{-1}\sigma_\varepsilon^2 |\Psi(e^{-i\lambda})|^2$, and therefore

$$(2\pi)^{-1}\sigma_\varepsilon^2 l^2 \le f(\lambda) \le (2\pi)^{-1}\sigma_\varepsilon^2 L^2. \tag{20}$$

Taking into account that $\gamma(0) = 1 = \int_{-\pi}^{\pi} f(\lambda)d\lambda$, we obtain (17). The inequality (18) is an immediate consequence of the following evident inequalities

$$|\gamma(k)| = \sigma_\varepsilon^2 \left| \sum_{j=0}^{\infty} \psi_j \psi_{j+|k|} \right| \le L^2 \sigma_\varepsilon^2 \rho^{-|k|} \sum_{j=0}^{\infty} \rho^{-2j} = \frac{L^2\sigma_\varepsilon^2\rho^2}{(\rho^2-1)}\rho^{-|k|}$$

and (17) (here we have used the bound on ψ_j established in (16)). The bounds on $\|\Gamma_d\|$ and $\|\Gamma_d^{-1}\|$ follow from the theorem on the eigenvalues of the Toeplitz forms [cf. Grenander and Szego (1984)]. In particular, we have

$$l^2\sigma_\varepsilon^2 \le \lambda_{\min}[\Gamma_d] \le \lambda_{\max}[\Gamma_d] \le L^2\sigma_\varepsilon^2,$$

where $\lambda_{\min}[\cdot]$ and $\lambda_{\max}[\cdot]$ denote the minimal and maximal eigenvalues of a matrix respectively. Applying (17) we obtain (19) which completes the proof. ■

A.1 An exponential inequality for sample covariances

Here we establish an exponential inequality on the deviation of sample covariances from their expectations. This result is basic for our future developments; furthermore, it is interesting in its own right.

Lemma 2 *Let $(X_t)_{t\in\mathbf{Z}} \in \mathcal{H}_\rho(l, L)$; then there exist absolute constants C_1 and C_2 such that for every integer k one has*

$$P\Big\{\Big|\frac{1}{n}\sum_{t=1}^{n} X_t X_{t+k} - \gamma(k)\Big| > \delta\Big\} \le$$

$$\begin{cases} \exp\Big(-\dfrac{\delta^2 n}{4C_1 M k_* r}\Big), & 0 \le \delta \le \Big(\dfrac{k_* r}{n}\Big)^{2/5}\Big(\dfrac{C_1^3 M^3}{C_2}\Big)^{1/5}, \\ \exp\Big(-\dfrac{1}{4}\Big[\dfrac{\delta n}{C_2 k_* r}\Big]^{1/3}\Big), & \delta \ge \Big(\dfrac{k_* r}{n}\Big)^{2/5}\Big(\dfrac{C_1^3 M^3}{C_2}\Big)^{1/5}, \end{cases} \tag{21}$$

where M and r are defined in (7) and $k_* = |k|$ whenever $k \neq 0$, and $k_* = 1$ whenever $k = 0$. The constants C_1 and C_2 are specified explicitly in the proof of the lemma.

Remark 4 To establish the result of the lemma we use the general exponential inequalities for weakly dependent random sequences found in Saulis and Statulevičus (1991). Several other exponential–type inequalities for weakly dependent random sequences appear already in the literature [cf., e.g., Doukhan (1994), Bosq (1996)], of which Bosq (Theorem 1.4, 1996) deals with conditions that are probably most akin to our set up. However, the machinery in Saulis and Statulevičus (1991) seems more suitable for our purposes, and leads to tighter bounds, in particular since we use the moderate deviations regime in (21).

Proof See Appendix C.1.

B Proofs of main results

B.1 Proof of Theorem 1

In the below proof $K_i, i = 1, 2, \ldots$ stand for absolute positive constants (unless otherwise specified), possibly different in different instances.

We first outline the main ideas in the proof. By straightforward algebra we have

$$\hat{\theta}^d - \theta^d = Q^{-1}\Big(-n^{-1}\theta^d + \frac{1}{n}\sum_{t=1}^{n} Z_t \sum_{j=d+1}^{\infty} \phi_j X_{t-j} + \frac{1}{n}\sum_{t=1}^{n} Z_t\varepsilon_t\Big), \tag{22}$$

where $Q \triangleq n^{-1}\sum_{t=1}^{n} Z_t Z_t' + n^{-1} I_d$. Thus, to prove a bound on the ℓ_2 distance between θ^d and $\hat{\theta}^d$, we must bound the norm of the matrix Q^{-1}, and of the vector multiplying it from the right in (22). The latter bound involves straightforward algebraic manipulations, therefore the real problem is to control the norm of Q^{-1}. The key idea here is the following. Partition the sample space into two sets. One set corresponds to the samples of $(X_t)_{t\in\mathbf{Z}}$, for which the elements of Q are uniformly 'close' to their expectations. For the complement of this set, $\|Q^{-1}\|$ does not grow faster then n. Exponential inequalities on the uniform convergence of sample means to their expectations, in the spirit of Lemma 2, ensure that the "bad" set essentially does not contribute to the overall bound. We shall now make these statements rigorous.

1^0. First, proceed to bound $\|Q^{-1}\|$, where $\|\cdot\|$ denotes the standard Euclidean matrix norm. Note that the i,j-entry Q_{ij} of the matrix Q with $i \neq j$ may be expressed as follows

$$\begin{aligned} Q_{ij} &= \frac{1}{n}\sum_{t=1}^{n} X_{t-i}X_{t-j} \\ &= \frac{1}{n}\sum_{\tau=1}^{n} X_\tau X_{\tau+j-i} - \frac{1}{n}\sum_{\tau=n-j+1}^{n} X_\tau X_{\tau+j-i} + \frac{1}{n}\sum_{\tau=1-j}^{0} X_\tau X_{\tau+j-i} \\ &\triangleq \hat{Q}_{ij} - W_{ij} + V_{ij}. \end{aligned}$$

This, in turn, may be written as

$$Q = \hat{Q} - W + V + n^{-1}I_d = \Gamma_d\left[I_d + \Gamma_d^{-1}\left(\tilde{Q} + n^{-1}I_d\right)\right] \tag{23}$$

where $\hat{Q} = (\hat{Q}_{ij})$, $W = (W_{ij})$, $V = (V_{ij})$, $i,j = 1,\ldots,d$, and $\tilde{Q} \triangleq V - W + \hat{Q} - \Gamma_d$. Observe that Γ_d is non–singular for every d (this follows from Lemma 1). Thus, the task of bounding $\|Q^{-1}\|$ is reduced to establishing a bound on the norm of $\left[I_d + \Gamma_d^{-1}\left(\tilde{Q} + n^{-1}I_d\right)\right]^{-1}\Gamma_d^{-1}$, where the only stochastic term is $\tilde{Q}$. The main idea is the following. Write

$$\tilde{Q} = (V - E[V]) - (W - E[W]) + (\hat{Q} - \Gamma_d) \quad ,$$

utilizing the fact that $E[V] = E[W]$. Note also that $E[\hat{Q}] = \Gamma_d$. In addition, due to Lemma 2 we can evaluate how close $\hat{Q}$ to Γ_d is. Now, the key to bounding $\|\tilde{Q}\|$, is to establish non–asymptotic exponential bounds on the probability that each one of the terms V, W, and $\hat{Q}$ deviate from their expectations.

Lemma 3 *Let $(X_t)_{t\in\mathbf{Z}} \in \mathcal{H}_\rho(l,L)$. For any fixed $i,j \in \{1,\ldots,d\}$ we have*

$$P\left\{|V_{ij} - EV_{ij}| > \delta\right\} \leq$$

$$\begin{cases} \exp\left(-\dfrac{\delta^2 n}{4C_1 d}\right), & 0 \leq \delta \leq \left[dn^{-1}C_1^2C_2^{-1}\right]^{1/3}, \\ \exp\left(-\dfrac{1}{4}\sqrt{\dfrac{\delta n}{C_2 d}}\right), & \delta \geq \left[dn^{-1}C_1^2C_2^{-1}\right]^{1/3}, \end{cases} \tag{24}$$

where C_1 and C_2 are as in Lemma 2. The same relations hold for W_{ij}.

Proof See Appendix C.2

2^0. Recall that by definition $EV_{ij} = EW_{ij}$, so that

$$\tilde{Q} = (V - E(V)) + (E(W) - W) + \hat{Q} - \Gamma_d \quad .$$

Applying the results of Lemma 2 and Lemma 3 we bound the norm of the matrix Q^{-1}. Let us fix $\kappa \in (0,1)$ and define the event

$$A_\kappa = \left\{ \omega \in \Omega : \max_{i,j=1,\dots,d} |\tilde{Q}_{ij}| \le C_\kappa \right\}, \tag{25}$$

where

$$C_\kappa = 6\sqrt{C_1 r M}\sqrt{\frac{d}{n}\ln\left(\frac{6d^2}{\kappa}\right)} \; . \tag{26}$$

Here, in and the sequel, Ω is the sample set of the underlying probability space $(\Omega, \mathcal{F}, P)$.

Lemma 4 *Let $(X_t)_{t\in\mathbf{Z}} \in \mathcal{H}_\rho(l,L)$ and for a fixed $\kappa \in (0,1)$ let d and n be such that*

$$d^{-1}n \ge \Big(36 C_1 r M \ln(6d^2/\kappa)\Big)^5, \tag{27}$$

and

$$n^{-1} + dC_\kappa \le \frac{1}{2}(l/L)^2. \tag{28}$$

Then $P(A_\kappa^c) \ge 1-\kappa$, and $\|\Gamma_d Q^{-1}\| \le 2$ if the event A_κ holds, and $\|Q^{-1}\| \le n$ otherwise.

Proof See Appendix C.3.

4^0. Now, recall for completeness (22)

$$\begin{aligned}
\hat{\theta}^d - \theta^d &= Q^{-1}\Big(-n^{-1}\theta^d + \frac{1}{n}\sum_{t=1}^n Z_t \sum_{j=d+1}^{\infty} \phi_j X_{t-j} + \frac{1}{n}\sum_{t=1}^n Z_t \varepsilon_t\Big) \\
&= Q^{-1}(\mathcal{I}_1 + \mathcal{I}_2 + \mathcal{I}_3) = Q^{-1}\mathcal{I}.
\end{aligned}$$

Having established a bound on $\|Q^{-1}\Gamma_d\|$ we proceed to bound $E\|\mathcal{I}\|$.

Lemma 5 *Let $(X_t)_{t\in\mathbf{Z}} \in \mathcal{H}_\rho(l,L)$. Then,*

$$\begin{aligned}
\|\mathcal{I}_1\| &\le \frac{1}{nl(\rho-1)} \\
\Big(E\|\mathcal{I}_2\|^4\Big)^{1/2} &\le \frac{K_1 d \rho^{-2d}}{l^2(\rho-1)^2} \\
\Big(E\|\mathcal{I}_3\|^4\Big)^{1/2} &\le \frac{K_2 d}{l^2 n},
\end{aligned}$$

where K_1 and K_2 are absolute constants.

Proof See Appendix C.4.

5^0. Now we complete proof of Theorem 1. We will proceed to bound $E\|\hat{\theta}^d - \theta^d\|^2$ by evaluating the expectation over two disjoint subsets corresponding to the events A_κ and A_κ^c. Let $\kappa = 6d^2n^{-6}$, A_κ be given by (25) with C_κ defined by (26) with κ in question. It can be immediately checked that under (8) conditions of the Lemma 4 hold. Thus we can write

$$\begin{aligned} E\left[\|\hat{\theta}^d - \theta^d\|^2 \mathbf{1}_{\{A_\kappa\}}\right] &\leq \|\Gamma_d^{-1}\|^2 E\left[\|\Gamma_d Q^{-1}\|^2 \, \|\mathcal{I}_1 + \mathcal{I}_2 + \mathcal{I}_3\|^2 \mathbf{1}_{\{A_\kappa\}}\right] \\ &\overset{(a)}{\leq} 4\|\Gamma_d^{-1}\|^2 E\left(\|\mathcal{I}_1 + \mathcal{I}_2 + \mathcal{I}_3\|^2 \mathbf{1}_{\{A_\kappa\}}\right) \\ &\leq 16\|\Gamma_d^{-1}\|^2 \left(\|\mathcal{I}_1\|^2 + E\|\mathcal{I}_2\|^2 + E\|\mathcal{I}_3\|^2\right) \\ &\overset{(b)}{\leq} K_3\|\Gamma_d^{-1}\|^2 l^{-2}\left(\frac{1}{n^2(\rho-1)^2} + \frac{d}{\rho^{2d}(\rho-1)^2} + \frac{d}{n}\right) \end{aligned}$$

where (a) follows from Lemma 4, and (b) follows from the bounds established in Lemma 5. Similarly, we have

$$\begin{aligned} E\left[\|\hat{\theta}^d - \theta^d\|^2 \mathbf{1}_{\{A_\kappa^c\}}\right] &\leq 4E\left[\|Q^{-1}\|^2\Big(\|\mathcal{I}_1\|^2 + \|\mathcal{I}_2\|^2 + \|\mathcal{I}_3\|^2\Big)\mathbf{1}_{\{A_\kappa^c\}}\right] \\ &\leq 4n^2\left[\|\mathcal{I}_1\|^2\mathbf{P}(A_\kappa^c) + \left(\sqrt{E\|\mathcal{I}_2\|^4} + \sqrt{E\|\mathcal{I}_3\|^4}\right)\sqrt{\mathbf{P}(A_\kappa^c)}\right] \\ &\leq K_4 n^2 l^{-2}\left[\frac{\kappa}{n^2(\rho-1)^2} + \frac{d\sqrt{\kappa}}{\rho^{2d}(\rho-1)^2} + \frac{\sqrt{\kappa}d}{n}\right]. \end{aligned}$$

Substituting expression for κ and combining the two bounds above we have

$$\left[E\|\hat{\theta}^d - \theta^d\|^2\right]^{1/2} \leq K_5\|\Gamma_d^{-1}\| l^{-1}\left(\frac{1}{n(\rho-1)} + \frac{\sqrt{d}}{\rho^d(\rho-1)} + \sqrt{\frac{d}{n}}\right), \tag{29}$$

whence

$$\begin{aligned} \left[E\|\hat{\phi} - \phi\|^2\right]^{1/2} &\leq \left[E\|\hat{\theta}^d - \theta^d\|^2\right]^{1/2} + \Big(\sum_{j=d+1}^{\infty} |\phi_j|^2\Big)^{1/2} \\ &\leq K_5\|\Gamma_d^{-1}\| l^{-1}\left(\frac{1}{n(\rho-1)} + \frac{\sqrt{d}}{\rho^d(\rho-1)} + \sqrt{\frac{d}{n}}\right) + \frac{1}{l\rho^d(\rho-1)}. \end{aligned}$$

Applying (19) completes the proof. ■

B.2 Proof of Theorem 2

Proof of the theorem rests upon the standard technique for deriving lower bounds in non-parametric estimation problems. In the proof below K_i, $i = 1, 2, \ldots$ denote positive constants depending on l and L only.

Let us fix a natural number N, and consider the following family $\mathcal{P}$ of the sequences $\phi = (\phi_1, \phi_2, \ldots)$: ϕ belongs to $\mathcal{P}$ if and only if

$$\phi_j = \begin{cases} \pm\beta\rho^{-N}, & j = 1, \ldots, N \\ 0, & \text{otherwise,} \end{cases}$$

where β is a positive number to be chosen. We complement $\mathcal{P}$ by the zero sequence $\phi^{(0)} = (0, 0, \ldots)$. It is evident that there exists a choice of constant K_1 (e.g., take $K_1 \leq \min\{1 - L^{-1}, l^{-1} - 1\}$), such that with the choice $\beta = K_1(1 - \rho^{-1})$ every $\phi \in \mathcal{P}$ defines a process $(X_t)_{t\in\mathbf{Z}}$ from $\mathcal{H}_\rho(l, L)$. In addition, cardinality of $\mathcal{P}$ is equal to $2^N + 1$. According to the Varshamov–Gilbert lemma [see, e.g., Korostelev and Tsybakov (1993, pp. 79)] one can choose a subfamily $\mathcal{P}' \subset \mathcal{P}$ so that any two distinct sequences ϕ', ϕ'' from $\mathcal{P}'$ differ by at least $N/16$ components, cardinality of $\mathcal{P}'$ is equal to $2^{\lfloor N/8 \rfloor} + 1$ and $\phi^{(0)} \in \mathcal{P}'$. Thus, for any ϕ', ϕ'' one has

$$\|\phi' - \phi''\| \geq K_2\sqrt{N}(1 - \rho^{-1})\rho^{-N} \triangleq s. \tag{30}$$

Let $\hat{\phi}_n$ be an arbitrary estimate of ϕ based on the data $\{X_t\}_{t=1}^n$; then

$$\sup_{(X_t)\in\mathcal{H}_\rho(m,L)} E\|\hat{\phi}_n - \phi\| \geq \sup_{\phi\in\mathcal{P}'} E\|\hat{\phi}_n - \phi\| \geq \frac{s}{2} \sup_{\phi\in\mathcal{P}'} P\left\{\|\hat{\phi}_n - \phi\| \geq s/2\right\}. \tag{31}$$

Now consider the problem of testing between $2^{\lfloor N/8 \rfloor} + 1$ hypotheses $H_j : \phi = \phi^{(j)}$ using observations $\{X_t\}_{t=1}^n$; here $\phi^{(j)}$, $j = 0, \ldots, 2^{\lfloor N/8 \rfloor}$ stand for the sequences from $\mathcal{P}'$. Define the decision rule $\tau : (X_1, \ldots, X_n) \mapsto \{0, \ldots, 2^{\lfloor N/8 \rfloor}\}$ as follows. Given the observations, we compute $\hat{\phi}_n$ and check to which of the sequences $\phi^{(j)} \in \mathcal{P}'$ it is closer in $\|\cdot\|$–distance. Then we have

$$\sup_{\phi\in\mathcal{P}'} P\left\{\|\hat{\phi} - \phi\| \geq s/2\right\} = \sup_{j=0,\ldots,2^{\lfloor N/8 \rfloor}} P\{\tau \neq j | H_j\}$$

and we should evaluate from below the probability of error under the decision rule τ. This can be done using the Fano inequality [see, e.g., Ibragimov and Has'minskii (1981, pp. 323)]. Let $g_j(y),\ j = 0, \ldots, 2^{\lfloor N/8 \rfloor}$ denote joint density of observations $X_1, \ldots, X_n$ under the hypothesis H_j. Denote by $\mathcal{K}(g_i, g_j)$ the Kullback–Leibler distance between the densities g_i and g_j. Then we have for $i \neq j$

$$\begin{aligned}
\mathcal{K}(g_i, g_j) &\leq \sup_{i,j} E_i \ln \frac{g_i(X_1, \ldots, X_n)}{g_j(X_1, \ldots, X_n)} \\
&\overset{(a)}{=} \sup_{i,j} E_i \left[-\frac{1}{2\sigma_\varepsilon^2} \sum_{t=1}^n \left(\left(X_t - \sum_{k=1}^N \phi_k^{(i)} X_{t-k} \right)^2 - \left(X_t - \sum_{k=1}^N \phi_k^{(j)} X_{t-k} \right)^2 \right) \right] \\
&\overset{(b)}{=} \sup_{i,j} \frac{n}{2\sigma_\varepsilon^2} E_i \left(\sum_{k=1}^N \left(\phi_k^{(i)} - \phi_k^{(j)} \right) X_{t-k} \right)^2 \\
&\leq \frac{n}{2\sigma_\varepsilon^2} \sup_{i,j} \sum_{k,l=1}^N \left(\phi_k^{(i)} - \phi_k^{(j)} \right) \left(\phi_l^{(i)} - \phi_l^{(j)} \right) E_i[X_{t-k} X_{t-l}] \\
&= \frac{n}{2\sigma_\varepsilon^2} \sup_{i,j} \left(\phi^{(i)} - \phi^{(j)} \right)' \Gamma_N^{(i)} \left(\phi^{(i)} - \phi^{(j)} \right),
\end{aligned}$$

where E_i denotes expectation with respect to the distribution related to the hypothesis H_i, and $\Gamma_N^{(i)} = \{E_i[X_{t-k}X_{t-l}]\}_{k,l=1}^N$ is the $N \times N$ covariance matrix under H_i. Here (a) follows from the fact that (X_t) is a Gaussian process, and (b) is obtained by taking expectation with respect to the density g_i. Using the bounds established in Lemma 1 on σ_ε^2 and on the maximal eigenvalue of the covariance matrix Γ_N (which are uniform over the class $\mathcal{H}_\rho(l, L)$ and N) we have

$$\mathcal{K}(g_i, g_j) \le \frac{nL^2}{2}(L/l)^2 \sup_{i,j} \|\phi^{(i)} - \phi^{(j)}\|^2 \le K_3 n N \rho^{-2N}.$$

Now set

$$N = \left\lfloor \frac{1}{2\ln\rho} \ln(K_4 n) \right\rfloor \ . \tag{32}$$

then due to the Fano inequality we can choose a constant K_4 so that under (12) probability of the error under τ will be at least, say, 1/4. Combining (30), (31) and (32) we come to the required statement. ■

B.3 Proof of Theorem 3

1^0. We have the following decomposition of the prediction error

$$\begin{aligned} X_{n+1} - \hat{X}_{n+1}^d &= \sum_{j=1}^d (\phi_j - \hat{\phi}_j) X_{n+1-j} + \sum_{j=d+1}^\infty \phi_j X_{n+1-j} + \varepsilon_{n+1} \\ &\triangleq \mathcal{E}_1 + \mathcal{E}_2 + \varepsilon_{n+1} \ . \end{aligned}$$

Therefore

$$E(X_{n+1} - \hat{X}_{n+1}^d)^2 = E(\mathcal{E}_1 + \mathcal{E}_2)^2 + \sigma_\varepsilon^2 \le 2\Big(E|\mathcal{E}_1|^2 + E|\mathcal{E}_2|^2\Big) + \sigma_\varepsilon^2 \ ,$$

where we have used the fact that ε_{n+1} is independent of X_t, for $t \le n$.

We first establish a bound on $E|\mathcal{E}_1|^2$. One clearly has

$$\begin{aligned} \mathcal{E}_1 &= \sum_{j=1}^d (\hat{\phi}_j - \phi_j) X_{n+1-j} \\ &= \sum_{j=1}^d (\hat{\phi}_j - \phi_j)\varepsilon_{n+1-j} + \sum_{k=1}^\infty \psi_k \sum_{j=1}^d (\hat{\phi}_j - \phi_j)\varepsilon_{n+1-j-k} \\ &\triangleq \eta_{n+1} + \sum_{k=1}^\infty \psi_k \eta_{n+1-k}, \end{aligned}$$

where the second equality follows from the MA(∞) representation of the process $(X_t)_{t\in\mathbf{Z}}$, and $\eta_t \triangleq \sum_{j=1}^d(\hat\phi_j-\phi_j)\varepsilon_{t-j}$. Thus, we have

$$
\begin{aligned}
E|\mathcal{E}_1|^2 &\le 4\Big[E|\eta_{n+1}|^2 + E\Big(\sum_{k=1}^d \psi_k\eta_{n+1-k}\Big)^2 + E\Big(\sum_{k=d+1}^\infty \psi_k\eta_{n+1-k}\Big)^2\Big]\\
&= 4\Big[E|\eta_{n+1}|^2 + \sum_{k,l=1}^d \psi_k\psi_l E[\eta_{n+1-k}\eta_{n+1-l}] + \sum_{k,l=d+1}^\infty \psi_k\psi_l E[\eta_{n+1-k}\eta_{n+1-l}]\Big]\\
&= 4(E|\eta_{n+1}|^2 + \mathcal{E}_{11} + \mathcal{E}_{12})\,.
\end{aligned}
$$

Let $\mathcal{F}_{-\infty}^n$ denote the σ–algebra on the common probability space $(\Omega,\mathcal{F},P)$ that is generated by the sequence $(\varepsilon_n,\varepsilon_{n-1},\ldots)$. We have

$$
\begin{aligned}
E|\eta_{n+1}|^2 &= E\sum_{i,j=1}^d(\hat\phi_i-\phi_i)(\hat\phi_j-\phi_j)\varepsilon_{n+1-i}\varepsilon_{n+1-j}\\
&= E\Big(\sum_{i,j=1}^d(\hat\phi_i-\phi_i)(\hat\phi_j-\phi_j)E\Big[\varepsilon_{n+1-i}\varepsilon_{n+1-j}|\mathcal{F}_{-\infty}^{\lfloor n/2\rfloor}\Big]\Big)\\
&= \sigma_\varepsilon^2 E\|\hat\theta^d-\theta^d\|^2, \qquad (33)
\end{aligned}
$$

where the second equality follows from from the fact that $\hat\phi$ is $\mathcal{F}_{-\infty}^{\lfloor n/2\rfloor}$-measurable and ε_{n+1-i}, $i=1,\ldots,d$ are independent of $\mathcal{F}_{-\infty}^{\lfloor n/2\rfloor}$ because $n/2>d$. Further, applying the same reasoning for $k,l=1,\ldots,d$ we obtain

$$
\begin{aligned}
E[\eta_{n+1-k}\eta_{n+1-l}] &= E\sum_{i,j=1}^d(\hat\phi_i-\phi_i)(\hat\phi_j-\phi_j)\varepsilon_{n+1-k-i}\varepsilon_{n+1-l-j}\\
&= E\Big(\sum_{i,j=1}^d(\hat\phi_i-\phi_i)(\hat\phi_j-\phi_j)E\Big[\varepsilon_{n+1-k-i}\varepsilon_{n+1-l-j}|\mathcal{F}_{-\infty}^{\lfloor n/2\rfloor}\Big]\Big)\\
&= \sigma_\varepsilon^2 E\sum_{i,j=1,\ i=l+j-k}^d(\hat\phi_i-\phi_i)(\hat\phi_j-\phi_j)\\
&\le K_1\sigma_\varepsilon^2 E\|\hat\theta^d-\theta^d\|^2.
\end{aligned}
$$

Here we again have used the fact that $\hat\phi$ is $\mathcal{F}_{-\infty}^{\lfloor n/2\rfloor}$-measurable, and $\varepsilon_{n+1-l-j}$, $i,l=1,\ldots,d$ are independent of $\mathcal{F}_{-\infty}^{\lfloor n/2\rfloor}$ because $n/4>d$. Therefore,

$$
\mathcal{E}_{11} \le K_1\sigma_\varepsilon^2 E\|\hat\theta^d-\theta^d\|^2\Big(\sum_{k=1}^d|\psi_k|\Big)^2 \le K_1\sigma_\varepsilon^2\frac{L^2}{(\rho-1)^2}E\|\hat\theta^d-\theta^d\|^2 \qquad (34)
$$

(here we have taken into account that $|\psi_k|\le L\rho^{-k}$).

To bound from above $\mathcal{E}_{12}$ note first that for $k\ge d+1$ by the Cauchy–Schwartz inequality

$$
\begin{aligned}
E|\eta_{n+1-k}|^2 = E\Big(\sum_{i=1}^d(\hat\phi_i-\phi_i)\varepsilon_{n+1-k-i}\Big)^2 &\le \sum_{i=1}^d E\Big[\|\hat\theta^d-\theta^d\|^2\varepsilon_{n+1-k-i}^2\Big]\\
&\le K_2 d\sigma_\varepsilon^2\Big(E\|\hat\theta^d-\theta^d\|^4\Big)^{1/2}.
\end{aligned}
$$

Thus, one has

$$\begin{aligned}\mathcal{E}_{12} &\le K_2 d\sigma_\varepsilon^2 \Big(E\|\hat{\theta}^d - \theta^d\|^4\Big)^{1/2} \Big(\sum_{k=d+1}^{\infty} |\psi_k|\Big)^2 \\ &\le K_2 d\sigma_\varepsilon^2 \Big(E\|\hat{\theta}^d - \theta^d\|^4\Big)^{1/2} \rho^{-2d} \frac{L^2}{(\rho-1)^2}. \qquad (35)\end{aligned}$$

Combining (35), (34) and (33) we come to the bound on $E|\mathcal{E}_1|^2$

$$E|\mathcal{E}_1|^2 \le K_3\sigma_\varepsilon^2 \Big[E\|\hat{\theta}^d - \theta^d\|^2 + \frac{d\rho^{-2d}L^2}{(\rho-1)^2}\Big(E\|\hat{\theta}^d - \theta^d\|^4\Big)^{1/2}\Big]. \qquad (36)$$

2^0. Our next step is to bound from above $E\|\hat{\theta}^d - \theta\|^4$. Choose $\kappa = 6d^2 n^{-10}$, and let A_κ be given by (25) with C_κ for κ in question. Write

$$E\|\hat{\theta}^d - \theta^d\|^4 = E\|\hat{\theta}^d - \theta^d\|^4 \mathbf{1}_{\{A_\kappa\}} + E\|\hat{\theta}^d - \theta^d\|^4 \mathbf{1}_{\{A_\kappa^c\}}.$$

It can be easily verified that under premise of the theorem the conditions of Lemma 4 hold. Therefore,

$$\begin{aligned}E\|\hat{\theta}^d - \theta^d\|^4 \mathbf{1}_{\{A_\kappa\}} &\le 2^4 \|\Gamma_d^{-1}\|^4 E\left(\|\mathcal{I}_1 + \mathcal{I}_2 + \mathcal{I}_3\|^4 \mathbf{1}_{\{A_\kappa\}}\right) \\ &\le K_4 \|\Gamma_d^{-1}\|^4 \left(\|\mathcal{I}_1\|^4 + E\|\mathcal{I}_2\|^4 + E\|\mathcal{I}_3\|^4\right).\end{aligned}$$

Applying Lemma 5 we obtain

$$E\|\hat{\theta}^d - \theta^d\|^4 \mathbf{1}_{\{A_\kappa\}} \le K_5 \|\Gamma_d^{-1}\|^4 l^{-4} \left(\frac{1}{n^4(\rho-1)^4} + \frac{d}{\rho^{4d}(\rho-1)^4} + \frac{d^2}{n^2}\right).$$

For the other term, involving the indicator of the event A_κ^c, the result follows the derivation in the proof of Theorem 1. In particular, we now require bounds in Lemma 5 to hold for $(E\|\mathcal{I}_j\|^8)^{1/2}$ for $j = 2, 3$. It is straightforward to extend the results of the lemma; the details are omitted. Thus, we obtain

$$E\|\hat{\theta}^d - \theta^d\|^4 \mathbf{1}_{\{A_\kappa^c\}} \le K_6 n^4 l^{-4} \left(\frac{\kappa}{n^4(\rho-1)^4} + \frac{d^2\sqrt{\kappa}}{\rho^{4d}(\rho-1)^4} + \frac{d^2\sqrt{\kappa}}{n^2}\right)$$

and finally, substituting $\kappa = 4d^2 n^{-10}$, and combining the above bounds we have

$$\Big(E\|\hat{\theta}^d - \theta^d\|^4\Big)^{1/2} \le K_7 (L^4/l^6) \left(\frac{1}{n^2(\rho-1)^2} + \frac{d}{\rho^{2d}(\rho-1)^2} + \frac{d}{n}\right).$$

Thus, it follows from (36) and (29) that

$$E|\mathcal{E}_1|^2 \le K_8 \sigma_\varepsilon^2 (L^4/l^6) \left(\frac{1}{n^2(\rho-1)^2} + \frac{d}{\rho^{2d}(\rho-1)^2} + \frac{d}{n}\right)\left(1 + \frac{d\rho^{-2d}L^2}{(\rho-1)^2}\right).$$

3^0. Now we complete proof of the theorem. We have the following upper bound on $E|\mathcal{E}_2|^2$

$$E|\mathcal{E}_2|^2 = E\bigg|\sum_{j=d+1}^{\infty} \phi_j X_{n+1-j}\bigg|^2 \le \bigg(\sum_{j=d+1}^{\infty} \phi_j\bigg)^2 \le l^{-2}\rho^{-2d}(\rho-1)^{-2},$$

where we have used the fact that $|\phi_k| \le l^{-1}\rho^{-k}$. Combining the above bounds on $E|\mathcal{E}_1|^2$ and $E|\mathcal{E}_2|^2$ we come to (14). This completes the proof of the theorem. ■

C Proofs of auxiliary results

C.1 Proof of Lemma 2

1^0. First observe that the process $(X_t)_{t\in\mathbf{Z}}$ is *strongly mixing*. We recall definition of the strong mixing condition [cf. Bradley (1986, pp. 169)]. For $-\infty \le s \le k \le \infty$ let $\mathcal{F}_s^k$ denote the σ-algebra generated by $(X_s, X_{s+1}, \dots, X_k)$. The process $(X_t)_{t\in\mathbf{Z}}$ is said to be strongly mixing if

$$\alpha_X(\tau) = \sup_{s\in\mathbf{Z}} \alpha(\mathcal{F}_{-\infty}^s, \mathcal{F}_{s+\tau}^\infty) \;\to\; 0, \;\text{ as }\; \tau\to\infty,$$

where

$$\alpha(\mathcal{F}_{-\infty}^s, \mathcal{F}_{s+\tau}^\infty) = \sup_{A\in\mathcal{F}_{-\infty}^s,\, B\in\mathcal{F}_{s+\tau}^\infty} |P(AB) - P(A)P(B)|.$$

Since $(X_t)_{t\in\mathbf{Z}}$ is Gaussian and stationary, we have $\alpha_X(\tau) = \alpha(\mathcal{F}_{-\infty}^0, \mathcal{F}_\tau^\infty)$. The strong mixing coefficient $\alpha_X(\tau)$ is bounded from above by the maximal correlation coefficient:

$$\alpha_X(\tau) \le \sup_{\zeta_1,\zeta_2} E(\zeta_1\zeta_2), \tag{37}$$

where the supremum in (37) is taken over all pairs of zero mean random variables (ζ_1, ζ_2) such that $\zeta_1 \in \mathcal{F}_{-\infty}^0, \zeta_2 \in \mathcal{F}_\tau^\infty$, and $E|\zeta_1|^2 = E|\zeta_2|^2 = 1$. Further, let $\mathcal{E}_{\tau-1}(f)$ denote the error of the best approximation of the spectral density $f(\lambda)$ by trigonometric polynomials of the degree $\le \tau-1$ on the interval $[-\pi,\pi]$ in the uniform norm. We have

$$\begin{aligned}\mathcal{E}_{\tau-1}(f) \le \frac{1}{\pi}\max_{\lambda\in[-\pi,\pi]}\bigg|\sum_{k=\tau}^{\infty}\gamma(k)\cos(\lambda k)\bigg| \;&\le\; \frac{1}{\pi}\sum_{k=\tau}^{\infty}|\gamma(k)| \\ &\le\; \frac{L^2\sigma_\varepsilon^2\rho^2}{\pi(\rho-1)^2}\rho^{-\tau},\end{aligned} \tag{38}$$

where the last inequality follows from (18). It is well–known [cf. Ibragimov and Rozanov (1978, pp. 146)] that for a stationary process with continuous and strictly positive spectral density the maximal correlation coefficient does not exceed $\Big[\min_{\lambda\in[-\pi,\pi]} f(\lambda)\Big]^{-1}\mathcal{E}_{\tau-1}(f)$. Taking into account (20) we obtain

$$\alpha_X(\tau) \le 2(L/l)^2\Big(\frac{\rho}{\rho-1}\Big)^2\rho^{-\tau}. \tag{39}$$

Now fix integer number k and define

$$U_{t,k} = \frac{1}{n} X_t X_{t+k} - \frac{\gamma(k)}{n}, \quad t \in \mathbf{Z}.$$

Without loss of generality we assume that k is non–negative integer number. Let $\mathcal{U}^s_{t,k}$ be the σ–algebra generated by $(U_{t,k}, U_{t+1,k}, \ldots, U_{s,k})$. Observe that $\mathcal{U}^t_{-\infty,k} \subseteq \mathcal{F}^{t+k}_{-\infty}$ and $\mathcal{U}^{\infty}_{t+\tau,k} \subseteq \mathcal{F}^{\infty}_{t+\tau}$. This implies that the process $(U_{t,k})_{t\in\mathbf{Z}}$ is also strongly mixing with the rate

$$\alpha_U(\tau) \le \alpha_X(\tau - k), \qquad \forall \tau > k.$$

For $\tau \le k$ we have the following trivial inequality $\alpha_U(\tau) \le 1$.

2^0. To complete proof of the lemma we need the following two auxiliary statements, adapted from Saulis and Statulevičus (1991, Theorem 4.17, Lemma 2.4).

Lemma 6 *Let $(Y_t)_{t\in\mathbf{Z}}$ be a strongly mixing random process, $S_n = \sum_{t=1}^n Y_t$, and $\mathrm{cum}_p(S_n)$ be the p-th order cumulant of the sum S_n. For $\nu > 0$ define the function*

$$\Lambda_n[\alpha_Y, \nu] = \max\Big\{1\,;\ \sum_{\tau=0}^{n} [\alpha_Y(\tau)]^{1/\nu}\Big\}.$$

If for some $\mu \ge 0, H > 0$

$$E|Y_t|^p \le (p!)^{\mu+1} H^p, \quad t = 1, \ldots, n, \quad p = 2, 3, \ldots,$$

then

$$|\mathrm{cum}_p(S_n)| \le 2^{p(1+\mu)+1} 12^{p-1} (p!)^{2+\mu} H^p \left\{\Lambda_n[\alpha_X, 2(p-1)]\right\}^{p-1} n.$$

For definition of the cumulants see, e.g., Brillinger (1975, pp. 19).

Lemma 7 *Let ξ be an arbitrary random variable with $E\xi = 0$. If there exist $\mu_1 \ge 0, H_1 > 0$ and $\Delta > 0$ such that*

$$|\mathrm{cum}_p(\xi)| \le \left(\frac{p!}{2}\right)^{1+\mu_1} \frac{H_1}{\Delta^{p-2}}, \quad p = 2, 3, \ldots,$$

then

$$P(|\xi| \ge x) \le \begin{cases} \exp\left\{-x^2/(4H_1)\right\}, & 0 \le x \le (H_1^{1+\mu_1}\Delta)^{1/(2\mu_1+1)} \\ \exp\left\{-(x\Delta)^{1/(1+\mu_1)}/4\right\}, & x \ge (H_1^{1+\mu_1}\Delta)^{1/(2\mu_1+1)}. \end{cases}$$

3^0. Using Lemma 6 we will derive upper bound on the cumulants of the sum $\sum_{t=1}^n U_{t,k}$, and then applying Lemma 7 we will obtain the required exponential inequality. First, we verify the conditions of Lemma 6 in order to apply it to the process $(U_{t,k})_{t\in\mathbf{Z}}$. Observe that for any natural p we have

$$\begin{aligned} E|U_{t,k}|^p &\le \frac{2^{p-1}}{n^p}\left(|\gamma(k)|^p + \left[E|X_t|^{2p}E|X_{t+k}|^{2p}\right]^{1/2}\right) \\ &\le \frac{2^{p-1}}{n^p}(1+p!2^p) \le \frac{2^{2p}p!}{n^p} \end{aligned}$$

where the second inequality follows from the fact that X_t is a Gaussian random variable, $|\gamma(k)| \le E|X_t|^2 = 1$, and $E|X_t|^{2p} \le p!\,2^p$. Further,

$$\begin{aligned} \Lambda_n[\alpha_U, 2(p-1)] &\le k + \sum_{\tau=k}^{n} [\alpha_X(\tau-k)]^{1/(2p-2)} \\ &\overset{(a)}{\le} k + \left(\frac{2L\rho}{l(\rho-1)}\right)^{2/(2p-2)} \sum_{\tau=0}^{n-k} \rho^{-\tau/(2p-2)} \\ &\le k + \left(\frac{2L\rho}{l(\rho-1)}\right)^{1/(p-1)} \frac{\rho^{1/(2p-2)}}{\rho^{1/(2p-2)}-1} \\ &\overset{(b)}{\le} k + \left(\frac{2L\rho}{l(\rho-1)}\right)^{1/(p-1)} \left(1+\frac{2p-2}{\ln\rho}\right) \end{aligned}$$

where (a) follows from the bound in (39), and (b) follows from the elementary inequality $\exp(x) - 1 \ge x$ for $x \ge 0$. Thus, one has

$$\begin{aligned} \{\Lambda_n[\alpha_U, 2(p-1)]\}^{p-1} &\le 2^{p-2}\left[k^{p-1} + \frac{2L\rho}{l(\rho-1)}(p-1)^{p-1}\left(1+\frac{2}{\ln\rho}\right)^{p-1}\right] \\ &\le 2^{p-2}k_*^{p-1}(p-1)!e^{p-1}\left(1+\frac{2}{\ln\rho}\right)^{p-1}\left(1+\frac{2L\rho}{l(\rho-1)}\right) \\ &\le (4e)^{p-1}p!(k_*r)^{p-1}M, \end{aligned}$$

where the inequality $(p-1)^{p-1} \le (p-1)!e^{p-1}$ has been used, and k_*, r, and M are defined in (7). Setting $\mu = 0$, and $H = 4n^{-1}$ we see that Lemma 6 applies for $(U_{t,k})_{t\in\mathbf{Z}}$, and thus

$$\left|\mathrm{cum}_p\Big(\sum_{t=1}^{n} U_{t,k}\Big)\right| \le 2^{3p+1}12^{2p-2}(p!)^3(k_*r)^{p-1}Mn^{-p+1}. \tag{40}$$

Now, to apply Lemma 7, put $\mu_1 = 2$, $H_1 = C_1Mk_*rn^{-1}$, and $\Delta = n(C_2k_*r)^{-1}$, where C_1, and C_2 are absolute constants ($C_1 = 2^{10}12^2$, $C_2 = 2^3 12^2$). It is immediately seen that the conditions of Lemma 7 hold for the parameters in question. Applying Lemma 7 completes the proof. ■

C.2 Proof of Lemma 3

The basis is the same argument as the one in Theorem 2. Recall the definition of V_{ij}

$$V_{ij} = \frac{1}{n} \sum_{\tau=1-j}^{0} X_\tau X_{\tau+j-i}$$

we have $EV_{ij} = jn^{-1}\gamma(j-i)$, whence $E|V_{ij}| \leq dn^{-1}$. Fix $i, j \in \{1, \ldots, d\}$ and define

$$U_t = \frac{1}{n} \left[X_t X_{t+j-i} - \gamma(j-i) \right],$$

then $V_{ij} - EV_{ij} = \sum_{t=1-j}^{0} U_t$. For any natural number p one has

$$\begin{aligned} E|U_t|^p &\leq \frac{2^{p-1}}{n^p} \left(|\gamma(j-i)|^p + \left[E|X_t|^{2p} E|X_{t+j-i}|^{2p} \right]^{1/2} \right) \\ &\leq p!\, 2^{2p} n^{-p} \end{aligned}$$

where the second inequality follows from the bound on $E|X_t|^{2p}$ established in step 4 of the proof of Theorem 2, and the fact that $|\gamma(k)| \leq 1, \; \forall\, k$.

Taking into account the strong mixing property of the sequence $(U_t)_{t\in\mathbf{Z}}$ and the fact that $\Lambda_j[\alpha_U, 2(p-1)] \leq j$ for $1 \leq j \leq d$, we can apply Lemma 6 with $\mu = 0$ and $H = 4n^{-1}$. Thus,

$$\left| \mathrm{cum}_p \left(\sum_{t=1-j}^{0} U_t \right) \right| \leq 2^{3p+1} 12^{p-1} (p!)^2 \frac{d^{p-1}}{n^{p-1}}. \tag{41}$$

It is immediately seen that the conditions of Lemma 7 hold with $\mu_1 = 1, \Delta = n(C_2 d)^{-1}$, and $H_1 = C_1 dn^{-1}$, and C_1 and C_2 may be chosen as in Theorem 2. The same argument is valid for W_{ij}. This completes the proof. ■

C.3 Proof of Lemma 4

(i). First we establish that $P(A_\kappa) \geq 1 - \kappa$. We have

$$\begin{aligned} P(A_\kappa^c) &\leq P\Big\{ \max_{i,j=1,\ldots,d} \Big(|V_{ij} - EV_{ij}| + |W_{ij} - EW_{ij}| + |(\hat{Q} - \Gamma_d)_{ij}| \Big) > C_\kappa \Big\} \\ &\leq P\Big\{ \max_{i,j=1,\ldots,d} |V_{ij} - EV_{ij}| > C_\kappa/3 \Big\} + P\Big\{ \max_{i,j=1,\ldots,d} |W_{ij} - EW_{ij}| > C_\kappa/3 \Big\} \\ &\quad + P\Big\{ \max_{i,j=1,\ldots,d} |(\hat{Q} - \Gamma_d)_{ij}| > C_\kappa/3 \Big\} \\ &\triangleq P_1 + P_2 + P_3. \end{aligned}$$

It can be easily verified that under the condition of (27), $C_\kappa \leq (d/n)^{2/5}$. Thus we may apply the results of Lemma 2 and Lemma 3 in the range of 'moderate' deviations. Note that P_3 can be bounded using the first inequality in (21) and the Toeplitz structure of the matrix $\hat{Q} - \Gamma_d$,

$$P_3 \le 2d \exp\Big\{-\frac{C_\kappa^2 n}{36 C_1 d r M}\Big\}.$$

The probabilities P_1 and P_2 are bounded, in turn, using the first inequality in (24)

$$P_i \le 2d^2 \exp\Big\{-\frac{C_\kappa^2 n}{36 C_1 d}\Big\}, \quad i = 1, 2.$$

Thus using the fact that $r \ge 1$ and $M \ge 1$, we have

$$P(A_\kappa^c) \le 6d^2 \exp\Big\{-\frac{C_\kappa^2 n}{36 C_1 d r M}\Big\}.$$

Now, it is straightforward to verify that the choice of C_κ is made so as to satisfy $P(A_\kappa^c) \le \kappa$.

(ii). Suppose that the event A_κ holds. Since $\tilde{Q}$ is a symmetric $d \times d$ matrix we have

$$\|\tilde{Q}\| = \lambda_{\max}(\tilde{Q}) \le \max_i \Big\{\sum_j |\tilde{Q}_{ij}|\Big\} \le dC_\kappa. \tag{42}$$

Therefore (23), and the definition of $\tilde{Q}$ together imply that

$$\|Q^{-1}\Gamma_d\| \le \frac{1}{1 - \|\Gamma_d^{-1}(\tilde{Q} + n^{-1} I_d)\|}, \tag{43}$$

provided that $\|\Gamma_d^{-1}(\tilde{Q} + n^{-1} I_d)\| < 1$. This condition will subsequently be verified. Taking into account (19) we have

$$\begin{aligned} \|\Gamma_d^{-1}(\tilde{Q} + n^{-1} I_d)\| &\le \|\Gamma_d^{-1}\| \left(n^{-1} + dC_\kappa\right) \\ &\le (L/l)^2 (n^{-1} + dC_\kappa) \le 1/2, \end{aligned} \tag{44}$$

where the last inequality follows from the condition imposed in (28). Thus, (44) along with (43) imply statement of the lemma for the case where the event A_κ holds.

(iii). Now consider the case of $\omega \in A_\kappa^c$. Independently of the event A_κ, the matrix Q is positive–definite, $\lambda_{\min}[Q] \ge n^{-1}$ and whence $\lambda_{\max}[Q^{-1}] \le n$. Since Q^{-1} is symmetric, we obtain immediately that $\|Q^{-1}\| \le n$. This completes the proof of the lemma. ■

C.4 Proof of Lemma 5

Upper bound on $\|\mathcal{I}_1\|$ follows immediately from (16):

$$\|\mathcal{I}_1\| = n^{-1}\|\theta^d\| = n^{-1}\Big(\sum_{j=1}^d |\phi_j|^2\Big)^{1/2} \le \frac{1}{nl(\rho - 1)}.$$

Let us denote the kth component of $\mathcal{I}_2$ as

$$\mathcal{I}_{2,k} \triangleq \frac{1}{n}\sum_{t=1}^{n} X_{t-k} \sum_{j=d+1}^{\infty} \phi_j X_{t-j}, \quad k = 1, 2, \ldots, d.$$

We have

$$E|\mathcal{I}_{2,k}|^4 = \sum_{j_1,\ldots,j_4=d+1}^{\infty} \phi_{j_1}\phi_{j_2}\phi_{j_3}\phi_{j_4} E\Big[\hat{\gamma}(k-j_1)\hat{\gamma}(k-j_2)\hat{\gamma}(k-j_3)\hat{\gamma}(k-j_4)\Big],$$

where $\hat{\gamma}(k-j) = n^{-1}\sum_{t=1}^{n} X_{t-k}X_{t-j}$. Applying repeatedly the Cauchy–Schwartz inequality and taking into account that $(X_t)_{t\in\mathbf{Z}}$ is Gaussian and stationary with $E|X_t|^2 = 1$, we obtain

$$E\Big[\hat{\gamma}(k-j_1)\hat{\gamma}(k-j_2)\hat{\gamma}(k-j_3)\hat{\gamma}(k-j_4)\Big] \le E|X_t|^8 \le 105.$$

Therefore

$$E|\mathcal{I}_{2,k}|^4 \le E|X_t|^8 \Big(\sum_{j=d+1}^{\infty} |\phi_j|\Big)^4 \le \frac{105}{\rho^{4d} l^4 (\rho-1)^4},$$

and thus

$$E\|\mathcal{I}_2\|^4 = E\sum_{k,l=1}^{d} |\mathcal{I}_{2,k}|^2 |\mathcal{I}_{2,l}|^2 \le \frac{105 d^2}{\rho^{4d} l^4 (\rho-1)^4}\ .$$

Now we derive an upper bound on $E\|\mathcal{I}_3\|^4$. Denote

$$\mathcal{I}_{3,k} \triangleq \frac{1}{n}\sum_{t=1}^{n} X_{t-k}\varepsilon_t = \frac{S_n}{n}, \qquad k = 1, 2, \ldots, d.$$

To bound $E|\mathcal{I}_{3,k}|^4 = n^{-4}|S_n|^4$ from above we note that $\{S_i, \mathcal{F}_{-\infty}^i, 1 \le i \le n\}$ is a martingale ($\mathcal{F}_{-\infty}^i = \sigma(\varepsilon_i, \varepsilon_{-1}, \ldots)$). Therefore due to Burkholder's inequality [see, e.g., Hall and Heyde (1980, pp. 23)] we have

$$E|S_n|^4 = E\left|\sum_{t=1}^{n} X_{t-k}\varepsilon_t\right|^4 \le K_1 E\left|\sum_{t=1}^{n} (X_{t-k}\varepsilon_t)^2\right|^2,$$

where K_1 is an absolute constant. Thus,

$$E|\mathcal{I}_{3,k}|^4 \le \frac{K_1}{n^4}\sum_{t,\tau=1}^{n} E\Big[X_{t-k}^2 X_{\tau-k}^2 \varepsilon_t^2 \varepsilon_\tau^2\Big] \le \frac{K_1}{n^2} E|X_t|^4 E|\varepsilon_t|^4 \le K_2 \frac{\sigma_\varepsilon^4}{n^2},$$

and finally

$$E\|\mathcal{I}_3\|^4 \le K_2 \frac{d^2\sigma_\varepsilon^4}{n^2} \le K_2 \frac{d^2}{l^4 n^2}\ ,$$

where the last inequality follows from (17). This completes the proof. ■

Acknowledgement: We would like to thank the referees for some helpful comments on a previous version of this manuscript.

References

[1] Akaike, H. (1970). Statistical predictor identification. *Ann. Inst. Statist. Math.* **21**, 243–247.

[2] Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automat. Control* **AC-19**, 716-723.

[3] An, H.-Z. Chen, Z.-G, C. and Hannan, E.J. (1982). Autocorrelation, autoregression and autoregressive approximation. *Ann. of Stat.* **10**, 926–936.

[4] Berk, K.N. (1974). Consistent autoregressive spectral estimates. *Ann. of Stat.* **2**, 489-502.

[5] Bhansali, R.J. (1981). Effects of not knowing the order of an autoregressive process on the mean-squared error of prediction. *J. Amer. Stat. Assoc.* **76**, 588-597.

[6] Bhansali, R.J.(1986). Asymptotically efficient selection of the order by the criterion autoregressive transfer function. *Ann. of Stat.* **14** 315–325.

[7] Bosq, D. (1996). *Nonparametric Statistics for Stochastic Processes.* Springer, New York.

[8] Bradley, R.C.(1986). Basic properties of strong mixing conditions. In *Dependence in Probability and Statistics*, Ed. E. Eberlein, M. Taqqu, 165–192, Birkhäuser, Boston.

[9] Brillinger, D.R.(1975) *Times Series: Data Analysis and Theory.* Holt, Rinehart and Winston, INC, New York.

[10] Bülmann, P. (1995). Moving average representations of autoregressive approximations. *Stoch. Proc. and Their Applic.* **60**, 331-342.

[11] Doukhan, P. (1994). *Mixing: Properties and Examples.* Springer, New York.

[12] Efromovich, S. (1998). Data–driven efficient estimation of the spectral density. *J. Amer. Stat. Assoc.* **92**, 762-769.

[13] Goldenshluger, A. (1998). Nonparametric estimation of transfer functions: rates of convergence and adaptation. *IEEE Trans. Inf. Theory* **44**, 644-658.

[14] Golubev, G. and Levit, B. (1996). Asymptotically efficient estimation for analytic distributions. *Math. Methods of Statistics* **5**, 357-368.

[15] Golubev, G. K., Levit, B. and Tsybakov A. (1996). Asymptotically efficient estimation of analytic functions in Gaussian noise. *Bernoulli* **2**, 167-181.

[16] Grenander, U. and Szegö, G. (1984). *Toeplitz Forms and Its Applications.* 2nd edition, Chelsea Publishing Company, New York.

[17] Gerencsér, L. (1992). AR(∞) estimation and non-parametric stochastic complexity. *IEEE Trans. on Info. The.* **38**, 1768-1778.

[18] Hall, P. and Heyde, C.C.(1980). *Martingale Limit Theory and Its Application.* Academic Press, New York.

[19] HANNAN, E.J. and KAVALIERIS, L. (1986). Regression, autoregression models. *J. of Time Ser. Anal.* **7**, 27-49.

[20] IBRAGIMOV, I.A. and HAS'MINSKII, R.Z. (1981). *Statistical Estimation*, New York, Springer–Verlag.

[21] IBRAGIMOV, I.A. and ROZANOV, Y.A. (1978). *Gaussian Random Processes*, Springer–Verlag, New York.

[22] KOROSTELEV, A.P. and TSYBAKOV, A.B. (1993). *Minimax Theory of Image Reconstruction*. Springer–Verlag, New York.

[23] PARZEN, E. (1974). Some recent advances in time series modelling. *IEEE Trans. Automat. Control* **AC-19**, 723–730.

[24] PARZEN, E. (1983). Autoregressive spectral estimation. In: *Time series in the frequency domain*, 221–247, Handbook of Statist., 3, North–Holland, Amsterdam.

[25] RISSANEN, J. (1983). Universal coding, information, prediction, and estimation. *IEEE Trans. Inf. Theory* **30** 629–636.

[26] RUDIN, W.(1974). *Real and Complex Analysis*. 2nd edition, McGraw–Hill, Inc., New York.

[27] SAULIS, L. and STATULEVIČUS, V.A.(1991). *Limit Theorems for Large Deviations*. Kluwer Academic Publishers, Dordrecht.

[28] SHIBATA, R. (1980). Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. *Ann. Statist.* **8** 147–164.

[29] SHIBATA, R. (1981). An optimal autoregressive spectral estimate. *Ann. Statist.* **9** 300–306.

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