Bootstrap Autoregressive Order Selection

Jürgen Franke, Jens-Peter Kreiss and Martin Moser

In this paper we deal with the problem of fitting an autoregression of order p to given data coming from a stationary autoregressive process with infinite order. The paper is mainly concerned with the selection of an appropriate order of the autoregressive model. Based on the so-called final prediction error (FPE) a bootstrap order selection can be proposed, because it turns out that one relevant expression occuring in the FPE is ready for the application of the bootstrap principle. Some asymptotic properties of the bootstrap order selection are proved. To carry through the bootstrap procedure an autoregression with increasing but non-stochastic order is fitted to the given data. The paper is concluded by some simulations.
Keywords: Autoregression; bootstrap; final prediction error; order selection.

1. Introduction

In this paper we deal with observations $X_1, \ldots, X_n$ which are realizations of an infinite order autoregressive model (AR(∞)-model) of the following type

$$X_t = \sum_{\ell=1}^{\infty} a_\ell X_{t-\ell} + \varepsilon_t, \qquad t \in \mathbb{Z} = \{0, \pm 1, \pm 2, \ldots\}.$$

The process $(\varepsilon_t : t \in \mathbb{Z})$ consists of independent and identically distributed (i.i.d.) real valued random variables on a probability space $(\Omega, \mathcal{A}, P)$ with (cumulative) distribution function F. Furthermore we assume

$$E\,\varepsilon_1 = 0, \qquad E\,\varepsilon_1^2 = \sigma^2 \in (0,\infty) \text{ and } E\,\varepsilon_1^4 < \infty .$$

The parameter $\mathbf{a} = (a_\ell : \ell \in \mathbb{N})$, $\mathbb{N} = \{1, 2, \ldots\}$, is absolutely summable and the generating function $1 - \sum_{\ell=1}^{\infty} a_\ell z^\ell$ has no zeros in the closed complex unit disk. More formally, $\mathbf{j}(\mathbf{a}) = (1, -a_1, -a_2, \ldots) \in \ell^1 = \{\mathbf{b} \in \mathbb{R}^{\mathbb{N}_0} : \|\mathbf{b}\|_1 = \sum_{\ell=0}^{\infty} |b_\ell| < \infty\}$, $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$, and $\mathbf{j}(\mathbf{a})$ is invertible in ℓ^1 with respect to the convolution $(\mathbf{b} * \mathbf{c})_\ell = \sum_{j=0}^{\ell} b_j c_{\ell-j}$. If $\boldsymbol{\alpha} = \mathbf{j}(\mathbf{a})^{-1}$ denotes the inverse of $\mathbf{j}(\mathbf{a})$ with respect to convolution, then the process $\mathbf{X} = (X_t)$ allows the following representation as an infinite order moving average process

(MA(∞)-process)

$$X_t = \sum_{\ell=0}^{\infty} \alpha_\ell \varepsilon_{t-\ell}\,, \qquad t \in \mathbb{Z}\,,$$

where $\alpha_0 = 1$ and the coefficients α_ℓ, $\ell \geq 1$, can be computed recursively from the convolution equation $\sum_{j=0}^{\ell} \alpha_j a_{\ell-j} = 0$.

The paper is devoted to the problem of fitting an autoregression of order p (AR(p)-model) to the given set of data $X_1, \ldots, X_n$. We start with a brief but careful study of the final prediction error (FPE). We obtain that one of two relevant terms is ready for an approximation through the bootstrap principle. In contrast to the construction of usual order selection procedures (e.g. FPE- or AIC-method), which heavily depend on the kind of the involved parameter estimator (mostly the usual Yule-Walker estimator or the closely related least squares (LS) estimator), the bootstrap approximation is open for other parameter estimates. This takes care of the fact that the more precise we can estimate the parameter of an AR(p)-approximation to the given data the higher we probably want to choose the definitive order to obtain a more precise fit. In this context we think of M-estimators or ML-estimators for non-normally distributed situations or so-called adaptive procedures.

The bootstrap procedure is based on a preliminary autoregressive approximation with a non-stochastic order $p_0(n)$ converging to infinity. The reader who is interested in a more complete theory for the bootstrap procedure in this AR(∞)-setup is referred to Kreiss (1988, 1997) and Bühlmann (1997).

The paper is concluded by some simulation results. There the properties of the bootstrap version of FPE are compared with the AIC-method.

2. An Approximation of the Final Prediction Error

In this section, we derive an approximation of the well-known FPE-criterion function which is of an appropriate form to apply the bootstrap. Here, we sometimes use heuristic arguments just to motivate this approximation. A rigorous formulation of the asymptotic properties of the order selection procedure, based on this approximation, is postponed to Section 5.

The optimal parameter of a fitted autoregression of order p is defined as

$$a(p) = \underset{c(p)\in \mathbb{R}^p}{\operatorname{argmin}}\; E\left(X_t - c(p)^T X_{t-1}(p)\right)^2$$

where $X_{t-1}(p) = (X_{t-1}, \ldots, X_{t-p})^T$. If $\gamma_h = EX_tX_{t+h}$, $h \in \mathbb{N}_0$, denote the autocovariances and

$$\Gamma(p) = \left(\gamma_{|i-j|} : i, j = 1, \ldots, p\right) \quad \text{and} \quad \gamma(p) = (\gamma_1, \ldots, \gamma_p)^T\,,$$

then $a(p)$ is given by the Yule–Walker equations

$$a(p) \;=\; \Gamma(p)^{-1}\gamma(p)\,. \tag{2.1}$$

We note that under our assumptions on the parameter $\mathbf{a}$ the $(p\times p)$–matrix $\Gamma(p)$ is always positive definite, $\Gamma(p)$ and $\Gamma(p)^{-1}$ are uniformly bounded in $p \in I\!N$ with respect to the operator norm $\|B\| := \sup\{\|Bx\|_2 : \|x\|_2 := (\sum x_i^2)^{\frac{1}{2}} = 1\}$ and for the autocovariance function $\boldsymbol{\gamma} = (\gamma_h : h \in I\!N_0)$ we have $\boldsymbol{\gamma} \in \ell^1$.
On the basis of the given observations $X_1, \ldots, X_n$ suppose that we have an estimator $\hat{a}(p) = (\hat{a}_1(p), \ldots, \hat{a}_p(p))^T$ of $a(p)$ for all p up to a maximal order $p(n)$. The FPE idea suggests to choose the order for the definitive autoregressive fit as

$$P_0(n) \;=\; \underset{1\le p\le p(n)}{\arg\min}\;\; E^{\mathbf{X}}E^{\mathbf{Y}}\left(Y_t - \hat{a}(p)^T Y_{t-1}(p)\right)^2 , \tag{2.2}$$

where $\mathbf{Y}$ is an independent copy of the time series $\mathbf{X}$. As the number of observations n tends to infinity, the maximal order $p(n)$ is also supposed to converge to infinity. Following Shibata (1980) we obtain

$$\begin{aligned} P_0(n) &= \underset{1\le p\le p(n)}{\arg\min}\;\; \left\{\sigma^2 + \|\mathbf{a} - a(p)\|_\Gamma^2 + E\,\|\hat{a}(p) - a(p)\|_{\Gamma(p)}^2\right\} \\ &= \underset{1\le p\le p(n)}{\arg\min}\;\; \left\{E\left(X_t - a(p)^T X_{t-1}(p)\right)^2 + E\,\|\hat{a}(p) - a(p)\|_{\Gamma(p)}^2\right\} , \end{aligned}$$

where $\Gamma = (\gamma_{|i-j|} : i,j \in I\!N)$, $\|x\|_B^2 = x^T B x$ and, for the sake of simplicity, $a(p)$ also denotes the ℓ^1-vector $(a_1(p), \ldots, a_p(p), 0, \ldots)$ filled up with zeros. Note that $P_0(n)$ is a deterministic but not computable quantity. Now the idea is to estimate both parts of the FPE. Let us start with the first expectation

$$E\left(X_t - a(p)^T X_{t-1}(p)\right)^2 \;=\; \gamma_0 - 2a(p)^T\gamma(p) + \|a(p)\|_{\Gamma(p)}^2 \,. \tag{2.3}$$

If we denote by $\hat{\boldsymbol{\gamma}}$ any consistent estimator of the autocovariance function $\boldsymbol{\gamma}$ (we will see later that $\hat{\boldsymbol{\gamma}}$ need not be the empirical autocovariances) this expression can be estimated by

$$\hat{\gamma}_0 - 2a(p)^T\hat{\gamma}(p) + \|a(p)\|_{\hat{\Gamma}(p)}^2 \,.$$

Next we intend to plug in a further estimate, namely estimators $\hat{a}(p)$ for $a(p)$. As the optimal parameters $a(p)$ correspond to the autocovariance function $\boldsymbol{\gamma}$ through the Yule–Walker equations (2.1), the same should hold for the estimators $\hat{a}(p)$ and $\hat{\boldsymbol{\gamma}}$ belonging to them, i.e.

$$\hat{\Gamma}(p)\hat{a}(p) \;=\; \hat{\gamma}(p)\,, \qquad 1 \le p \le p(n)\,. \tag{2.4}$$

Obviously, $\hat{\boldsymbol{\gamma}}$ needs only to be known up to lag $p(n)$. For ease of notation we do not explicitly indicate the dependence of the estimators on the number n of observations.
This second substitution introduces a systematic bias which, using (2.4), may be calculated as

$$\begin{aligned} &\left(\hat{\gamma}_0 - 2a(p)^T\hat{\gamma}(p) + \|a(p)\|_{\hat{\Gamma}(p)}^2\right) - \left(\hat{\gamma}_0 - 2\hat{a}(p)^T\hat{\gamma}(p) + \|\hat{a}(p)\|_{\hat{\Gamma}(p)}^2\right) \\ &\quad= -2\left(a(p) - \hat{a}(p)\right)^T\hat{\gamma}(p) + \left(a(p) - \hat{a}(p)\right)^T\hat{\Gamma}(p)\left(a(p) + \hat{a}(p)\right) \\ &\quad= \|a(p) - \hat{a}(p)\|_{\hat{\Gamma}(p)}^2 \,. \end{aligned}$$

Because of this a reasonable approximation of the argument of $P_0(n)$ is given by

$$\hat{\gamma}_0 - \hat{a}(p)^T\hat{\gamma}(p) + 2 \cdot E\,\|\hat{a}(p) - a(p)\|^2_{\Gamma(p)}\,, \tag{2.5}$$

where the expectation is ready for an approximation through the bootstrap, which we will discuss in detail in the next but one section. Note that we have so far not made any assumptions on the estimates $\hat{a}(p)$ and $\hat{\boldsymbol{\gamma}}$ except (2.4).

Finally we need a further approximation of the expectation in (2.5) in order to be able to evaluate some asymptotic properties of the bootstrap order selection. To this end observe that

$$\|\hat{a}(p) - a(p)\|^2_{\hat{\Gamma}(p)} = \left\|\hat{\Gamma}(p)^{-1}\left(\hat{\gamma}(p) - \hat{\Gamma}(p)a(p)\right)\right\|^2_{\hat{\Gamma}(p)} = \left\|\hat{\gamma}(p) - \hat{\Gamma}(p)a(p)\right\|^2_{\hat{\Gamma}(p)^{-1}}.$$

We will now make use of the following approximation of the argument of $P_0(n)$

$$\hat{\gamma}_0 - \hat{a}(p)^T\hat{\gamma}(p) + 2 \cdot E\left\|\hat{\gamma}(p) - \hat{\Gamma}(p)a(p)\right\|^2_{\Gamma(p)^{-1}}. \tag{2.6}$$

The construction of a bootstrap version of the following theoretical order selection procedure

$$P_1(n) := \operatorname*{arg\,min}_{1\le p\le p(n)} \left\{\hat{\gamma}_0 - \hat{a}(p)^T\hat{\gamma}(p) + 2 \cdot S_n(p)\right\} \tag{2.7}$$

where

$$\begin{aligned} S_n(p) &= E\left\|\hat{\gamma}(p) - \hat{\Gamma}(p)a(p)\right\|^2_{\Gamma(p)^{-1}} \\ &= E\left\|(\hat{\gamma}(p) - \gamma(p)) - \left(\hat{\Gamma}(p) - \Gamma(p)\right)a(p)\right\|^2_{\Gamma(p)^{-1}} \end{aligned} \tag{2.8}$$

is exactly the goal of the next but one section. Of course $P_1(n)$ is closely related to

$$P_1'(n) := \operatorname*{arg\,min}_{1\le p\le p(n)} \left\{\hat{\gamma}_0 - \hat{a}(p)^T\hat{\gamma}(p) + 2 \cdot E\,\|\hat{a}(p) - a(p)\|^2_{\Gamma(p)}\right\}. \tag{2.9}$$

3. Estimators of Prediction Coefficients

In this section we want to present briefly some estimators for the autocovariance function $\boldsymbol{\gamma}$ or the parameter value $a(p)$ of an autoregressive fit of order p which we have in mind. The easiest situation is to use the empirical autocovariances

$$\tilde{\gamma}_h = \frac{1}{n}\sum_{t=1}^{n-h} X_t X_{t+h}\,, \qquad h \in I\!N_0,$$

(or, asymptotically equivalent, their centered version $\tilde{\gamma}_h^c = \frac{1}{n}\sum_{t=1}^{n-h}(X_t - \frac{1}{n}\sum_{s=1}^{n} X_s)(X_{t+h} - \frac{1}{n}\sum_{s=1}^{n} X_s)$), to which belong the well-known Yule-Walker parameter estimators

$$\tilde{a}(p) = \tilde{\Gamma}(p)^{-1}\tilde{\gamma}(p)\,.$$

In this text we always equip empirical autocovariances and the corresponding Yule-Walker parameter estimators with a tilde. In contrast to these estimators we propose the following alternative. Fit in a first step an autoregression of (high) order $p_M \geq p(n)$ to the given data and compute M- or ML-parameter estimators, i.e. solutions of

$$\Psi_n\left(c_1, \ldots, c_{p_M}\right) = \sum_{t=p_M+1}^{n} \psi\left(X_t - \sum_{\ell=1}^{p_M} c_\ell X_{t-\ell}\right) X_{t-1}(p_M) \equiv 0\,, \tag{3.1}$$

where $\psi : I\!R \to I\!R$ denotes a suitable score function. We do not intend to discuss at this place the problem of finding solutions of (3.1). If $p_M = p_M(n)$ converges to infinity with an appropriate rate and if ψ satisfies some regularity conditions it is possible to find a solution of (3.1), denoted by $\hat{a}^M = \hat{a}^M(p_M)$, which is consistent for $\mathbf{a}$, cf. Kreiss (1988) and Moser (1997). As $\mathbf{j}(\mathbf{a})$ is invertible in ℓ^1 and the set of invertible sequences in ℓ^1 is open in ℓ^1, we may assume that $\mathbf{j}\left(\hat{a}^M\right)$ is invertible as well.

Denote the autocorrelation function belonging to an autoregressive process of order p_M with parameter $\hat{a}^M$ by $\hat{\mathbf{r}}$. As the autocovariance function $\mathbf{r}$ is a continuous function of the parameter $\mathbf{a}$ (with respect to $\|\cdot\|_1$), $\hat{\mathbf{r}}$ will be a consistent estimator for the theoretical autocorrelation function $\mathbf{r} = \frac{1}{\gamma_0}\boldsymbol{\gamma}$. The estimate $\hat{\mathbf{r}}$ coincides with the empirical autocorrelation function up to lag p_M if and only if $\hat{a}^M$ is the Yule–Walker estimate. $\hat{\mathbf{r}}$ may be computed using the MA(∞)–representation of the AR–process with parameter $\hat{a}^M$. As only the components of $\hat{\mathbf{r}}$ up to lag $p(n) \leq p_M$ are needed, an easier approach is to solve

$$\hat{R}(p_M)\,\hat{a}^M = \hat{r}(p_M)\,,$$

or equivalently

$$\hat{C}(p_M - 1)\begin{pmatrix} \hat{r}_1 \\ \vdots \\ \hat{r}_{p_M-1} \end{pmatrix} = -\begin{pmatrix} \hat{a}_1^M \\ \vdots \\ \hat{a}_{p_M-1}^M \end{pmatrix}, \qquad \hat{r}_{p_M} = \hat{a}_{p_M}^M + \sum_{\ell=1}^{p_M-1} \hat{a}_{p_M-\ell}^M \hat{r}_\ell\,,$$

where $\hat{C}(p_M - 1) = \left(\hat{a}_{i+j}^M + \hat{a}_{i-j}^M : 1 \leq i, j \leq p_M - 1\right)$, $\hat{a}_0^M = -1$ and $\hat{a}_k^M = 0$ if $k < 0$ or $k > p_M$. As $\mathbf{j}(\hat{a}^M)$ is supposed to be invertible in ℓ^1, the matrix $\hat{C}(p_M - 1)$ will be invertible, too.

Based on the autocorrelation estimates $\hat{\mathbf{r}}$, we may calculate new estimates of $a(p)$ using the Yule-Walker equations:

$$\hat{a}(p) = \hat{R}(p)^{-1}\hat{r}(p), \qquad p = 1, 2, \ldots, p(n)\,. \tag{3.2}$$

Why do we introduce such estimators? It is known that M-estimators $\hat{a}^M$ are more efficient for $\mathbf{a}$ if the innovations ε_t are not normally distributed, see e.g. Martin (1983) or Kreiss (1988). In particular, this is true if the distribution of the innovations has a Lebesgue density f and we take $\psi = -f'/f$, i.e. if we use ML-estimates. The gain in efficiency carries over to the estimators $\hat{\mathbf{r}}$ and $\hat{a}(p)$, which are smooth functions of $\hat{a}^M = \hat{a}^M(p_M)$.

It is not possible to use the M-estimators $\hat{a}^M(p)$ directly as estimators for $a(p)$, because, in general, they are not even consistent. Therefore, we have to use the detour of calculating $\hat{\mathbf{r}}$ from $\hat{a}^M(p_M)$ and then define $\hat{a}(p)$ as in (3.2) to get robust and consistent estimates of $a(p)$ for $p \leq p_M$. For a different approach where the quadratic loss function in (2.2) is replaced by some loss function L_ψ and where the optimal parameters may be estimated directly by the M-estimators, see Behrens (1990).

To avoid too much technical details, we consider in the following two sections only the easiest case where $\hat{r}_h = \tilde{r}_h$, $h \leq p(n)$, are the empirical autocorrelations and $\hat{a}(p) = \tilde{a}(p)$ are the Yule-Walker estimates. A theoretical investigation of the asymptotic properties of our bootstrap order selection procedure when $\hat{\mathbf{r}}$ corresponds to some M-estimator $\hat{a}^M(p_M)$ is considerably more involved and will be the subject of a forthcoming paper, see also Moser (1997).

4. Bootstrap Order Selection

Let us first briefly introduce the bootstrap principle for AR(∞)-processes which will be applied in the following. For a fuller account the interested reader is referred to Kreiss (1988, 1997) and Bühlmann (1997).

Given the observations $X_1, \ldots, X_n$ we fit an autoregression of "large" order $p_0 = p_0(n) \geq p(n)$ and compute approximate innovations

$$\varepsilon_t(\tilde{a}(p_0)) \;=\; X_t - \tilde{a}(p_0)^T X_{t-1}(p_0)\,, \qquad t = p_0 + 1, \ldots, n\,,$$

with empirical and centered (around mean 0) empirical (cumulative) distribution functions $\tilde{F}_n$, $\tilde{F}_n^c$, respectively.

Now suppose that the process $(\varepsilon_t^* : t \in \mathbb{Z})$ consists of i.i.d. random variables with distribution function $\tilde{F}_n^c$. This ensures $E^* \varepsilon_t^* = 0$, where E^* denotes the conditional expectation $E[\cdot|X_1, \ldots, X_n]$. It is well-known known that $1 - \sum_{\ell=1}^{p_0} \tilde{a}_\ell(p_0) z^\ell$ has no zeros in the closed unit disk and therefore $\mathbf{j}(\tilde{a}(p_0))$ has an inverse $\tilde{\boldsymbol{\alpha}}(p_0)$ in ℓ^1 with respect to convolution; cf. Brockwell and Davis (1991), p. 240. Hence we may define the bootstrap process $(X_t^* : t \in \mathbb{Z})$ as an autoregression of order p_0 with coefficients $\mathbf{a}^* = \tilde{a}(p_0)$ and white noise process $(\varepsilon_t^* : t \in \mathbb{Z})$, i.e.

$$X_t^* \;=\; \sum_{\ell=1}^{p_0} \tilde{a}_\ell(p_0) X_{t-\ell}^* + \varepsilon_t^* \;=\; \sum_{\ell=0}^{\infty} \tilde{\alpha}_\ell(p_0)\varepsilon_{t-\ell}^*\,, \qquad t \in \mathbb{Z}\,.$$

For later reference we note some asymptotic properties of the bootstrap construction. Assuming $p_0(n) \to \infty$ and $p_0(n)^4/n \to 0$ we have from Kreiss (1997) $E^* \varepsilon^{*k} \to E\,\varepsilon^k$ in probability, $k = 2, 4$, (Proposition 3.1) and $\|\boldsymbol{\alpha} - \tilde{\boldsymbol{\alpha}}(p_0)\|_1 \to 0$ in probability (Lemma 8.2 and 8.3). As the autocovariance function $\boldsymbol{\gamma}$ is a continuous function of $E\,\varepsilon_0^2 \in \mathbb{R}$ and $\boldsymbol{\alpha} \in \ell^1$ this implies $\|\boldsymbol{\gamma} - \boldsymbol{\gamma}^*\|_1 \to 0$ in probability for the autocovariance function $\boldsymbol{\gamma}^*$ with components $\gamma_h^* = E^* X_t^* X_{t+h}^*$ of the bootstrap autoregressive process.

Of course $\boldsymbol{\gamma}^*$ is closely related to the empirical autocovariance function $\tilde{\boldsymbol{\gamma}}$. In fact, the corresponding autocorrelation functions $\mathbf{r}^* = \frac{1}{\gamma_0^*}\boldsymbol{\gamma}^*$ and $\tilde{\mathbf{r}} = \frac{1}{\tilde{\gamma}_0}\tilde{\boldsymbol{\gamma}}$ coincide up to lag p_0, where both γ_0^* and $\tilde{\gamma}_0$ converge to γ_0 in probability. In particular, we have

$$a^*(p) \;=\; \Gamma^*(p)^{-1}\gamma^*(p) \;=\; \tilde{\Gamma}(p)^{-1}\tilde{\gamma}(p) \;=\; \tilde{a}(p)\,, \qquad 1 \le p \le p_0\,.$$

Let $\tilde{\boldsymbol{\gamma}}^*$ and $\tilde{a}^*(p)$ be exactly defined as $\tilde{\boldsymbol{\gamma}}$ and $\tilde{a}(p)$, with $(X_1,\ldots,X_n)$ replaced by $(X_1^*,\ldots,X_n^*)$, the bootstrap observations.
We propose to replace $E\,\|\tilde{a}(p) - a(p)\|^2_{\Gamma(p)}$ in the definition of the order selection $P_1'(n)$ (cf. (2.9)) by its bootstrap approximation

$$E^*\,\|\tilde{a}^*(p) - a^*(p)\|^2_{\Gamma^*(p)} \;=\; E^*\,\|\tilde{a}^*(p) - \tilde{a}(p)\|^2_{\Gamma^*(p)}\;.$$

To avoid technical problems, we will work with the order selection $P_1(n)$ instead of $P_1'(n)$, so we will use the bootstrap analogon

$$S_n^*(p) = E^*\left\|\tilde{\gamma}^*(p) - \tilde{\Gamma}^*(p)\tilde{a}(p)\right\|^2_{\Gamma^*(p)^{-1}}\;. \tag{4.1}$$

of $S_n(p)$, as we already mentioned in Section 2, cf. (2.7).

Hence we define the bootstrap order selection as

$$P_B(n) \;:=\; \operatorname*{argmin}_{1\le p\le p(n)} \;\left\{\tilde{\gamma}_0 - \tilde{a}(p)^T\tilde{\gamma}(p) + 2\cdot S_n^*\,(p)\right\}\;. \tag{4.2}$$

We remark that the whole procedure resulting in the order selection $P_B(n)$ can be done with general autocovariance estimates $\hat{\boldsymbol{\gamma}}$ and the corresponding sample prediction coefficients $\hat{a}(p)$, given by (3.2). We restrict ourselves to the sample autocovariances $\tilde{\boldsymbol{\gamma}}$ and the Yule-Walker estimates $\tilde{a}(p)$ only to simplify the proofs. One of our main results is as follows.

Theorem 4.1 : *Let $\{p(n) : n \in I\!N\}$ and $\{p_0(n) : n \in I\!N\}$ be two sequences of integers with $p(n) \le p_0(n)$ for all $n \in I\!N$ and $p(n) \to \infty$, $p_0(n)^4/n \to 0$ as $n \to \infty$. Then we have for $S_n(p)$, $S_n^*(p)$ defined in (2.8) and (4.1)*

$$\max_{1\le p\le p(n)} \left\{\frac{n}{p}\,|S_n^*(p) \;-\; S_n(p)|\right\} \;=\; o_P(1)\,.$$

All proofs are collected in Section 7.

As will be seen in the proof of Theorem 5.1, $\frac{n}{p}S_n(p) \to \sigma^2$ if p is "large". Hence Theorem 4.1 basically maintains that the difference between $S_n(p)$ and its bootstrap approximation $S_n^*(p)$ tends faster to 0 than $S_n(p)$ itself, i.e. $S_n(p)^{-1}\,|S_n^*(p) - S_n(p)| = o_P(1)$, uniformly in all "large" p. If p is "small", then $\tilde{\gamma}_0 - \tilde{a}(p)^T\tilde{\gamma}(p)$ will be the dominating term in $\tilde{\gamma}_0 - \tilde{a}(p)^T\tilde{\gamma}(p) - 2S_n(p)$, the expression minimized by $P_1(n)$, cf. (2.7). In this case the relative difference $S_n(p)^{-1}\,|S_n^*(p) - S_n(p)|$ will be of secondary importance as long as the absolute difference tends to 0 fast enough.

5. Asymptotic Properties of the Bootstrap Order Selection

In this section we deal with some asymptotic properties of the proposed bootstrap order selection $P_B(n)$ defined by (4.2):

$$P_B(n) \;=\; \underset{1\le p\le p(n)}{\arg\min} \;\; \left\{ \tilde{\gamma}_0 - \tilde{a}(p)^T \tilde{\gamma}(p) + 2\cdot S_n^*(p) \right\} .$$

The first part of the criterion function can be written in the more familiar way

$$\tilde{\gamma}_0 - \tilde{a}(p)^T \tilde{\gamma}(p) \;=\; \frac{1}{n} \sum_{t=p_0(n)+1}^{n} \left(X_t - \tilde{a}(p)^T X_{t-1}(p) \right)^2 \;=:\; \tilde{\sigma}_n^2(p) \,,$$

if we assume that, for the sake of simplicity, we slightly modify the definitions of Section 3 to

$$\tilde{\gamma}_h \;:=\; \frac{1}{n} \sum_{t=p(n)+1}^{n} X_t X_{t-h} \,, \qquad h = 0,1,2,...,p(n) \tag{5.1}$$

and

$$\tilde{a}(p) \;:=\; \left[\frac{1}{n} \sum_{t=p(n)+1}^{n} X_{t-1}(p) X_{t-1}(p)^T \right]^{-1} (\tilde{\gamma}_1, \ldots, \tilde{\gamma}_p)^T \;\equiv\; \tilde{\Gamma}(p)^{-1} \tilde{\gamma}(p) \,, \tag{5.2}$$

which are essentially the usual Yule-Walker estimators up to asymptotically negligible terms.

Now we state the main result of this section. Again, the proof is deferred to Section 7.

Theorem 5.1 : *Under the assumptions of Theorem 4.1 we have*

$$\max_{1\le p\le p(n)} \frac{\left| S_n^*(p) - \frac{p}{n}\sigma^2 \right|}{\frac{p}{n} + \frac{\sigma^2(p)-\sigma^2}{\sigma^2}} \;=\; o_P(1) \;,$$

where $\sigma^2(p) = E\,\left(X_t - a(p)^T X_{t-1}(p) \right)^2$.

¿From this result we can derive an interesting property of the bootstrap order selection, observing that

$$\max_p \frac{\frac{p}{n}\left|\sigma^2 - \sigma^2(p)\right|}{\frac{p}{n} + \frac{\sigma^2(p)-\sigma^2}{\sigma^2}} \le \max_p \frac{\frac{p}{n}\left|\sigma^2 - \sigma^2(p)\right|}{\frac{\sigma^2(p)-\sigma^2}{\sigma^2}} = \frac{p(n)}{n}\sigma^2 = o(1)$$

and

$$\max_p \frac{\frac{p}{n}\left|\tilde{\sigma}_n^2(p) - \sigma^2(p)\right|}{\frac{p}{n} + \frac{\sigma^2(p)-\sigma^2}{\sigma^2}} \;\le\; \max_p \left| \tilde{\sigma}_n^2(p) - \sigma^2(p) \right|$$

$$\le\; \max_p \left| \tilde{\sigma}_n^2(p) - \frac{1}{n} \sum_{t=p(n)+1}^{n} \left(X_t - a(p)^T X_{t-1}(p) \right)^2 \right|$$

$$
\begin{aligned}
&+ \max_p \left| \frac{1}{n} \sum_{t=p(n)+1}^{n} \left(X_t - a(p)^T X_{t-1}(p) \right)^2 - \sigma^2(p) \right| \\
= \;& \max_p \|\tilde{a}(p) - a(p)\|^2_{\Gamma(p)} + \max_p \left| \frac{1}{n} \sum_t \left(X_t - a(p)^T X_{t-1}(p) \right)^2 - \sigma^2(p) \right| \\
= \;& o_P(1)
\end{aligned}
$$

as the first expression is $o_P(1)$ by (7.6) and (7.9) and the second expression by an application of Lemma 7.1 to the process $Y_t = X_t - a(p)^T X_{t-1}(p)$.

This together with Theorem 5.1 and with Theorem 7.4.7 of Deistler and Hannan (1988) implies that we have uniformly in $p \in \{1, ..., p(n)\}$

$$
\begin{aligned}
& \log \left\{ \tilde{\sigma}_n^2(p) + 2 \cdot S_n^*(p) \right\} \\
= \;& \log \tilde{\sigma}_n^2(p) + \log \left(1 + 2 \cdot \frac{S_n^*(p)}{\tilde{\sigma}_n^2(p)} \right) \\
= \;& \log \tilde{\sigma}_n^2(p) + \log \left(1 + \frac{p}{n} \left[\frac{2}{\tilde{\sigma}_n^2(p)} \left\{ \frac{n}{p} S_n^*(p) - \tilde{\sigma}_n^2(p) \right\} + 2 \right] \right) \\
= \;& \log \tilde{\sigma}_n^2(p) + 2\frac{p}{n} + \frac{p}{n} \left[\frac{2}{\tilde{\sigma}_n^2(p)} \left\{ \frac{n}{p} S_n^*(p) - \tilde{\sigma}_n^2(p) \right\} + o_P(1) \right] \\
= \;& \log \dot{\sigma}_n^2 + \left\{ \frac{p}{n} + \frac{\sigma^2(p) - \sigma^2}{\sigma^2} \right\} (1 + o_P(1)) \\
& + \frac{p}{n} \left[\frac{2}{\tilde{\sigma}_n^2(p)} \left\{ \frac{n}{p} S_n^*(p) - \tilde{\sigma}_n^2(p) \right\} + o_P(1) \right] \\
= \;& \log \dot{\sigma}_n^2 + \left\{ \frac{p}{n} + \frac{\sigma^2(p) - \sigma^2}{\sigma^2} \right\} (1 + o_P(1)) \\
& + \frac{2}{\tilde{\sigma}_n^2(p)} \cdot \frac{S_n^*(p) - p/n \tilde{\sigma}_n^2(p)}{p/n + \frac{\sigma^2(p) - \sigma^2}{\sigma^2}} \left\{ \frac{p}{n} + \frac{\sigma^2(p) - \sigma^2}{\sigma^2} \right\} \\
= \;& \log \dot{\sigma}_n^2 + \left\{ \frac{p}{n} + \frac{\sigma^2(p) - \sigma^2}{\sigma^2} \right\} (1 + o_P(1)) \;, \qquad (5.3)
\end{aligned}
$$

where $\dot{\sigma}_n^2$ is defined in Deistler and Hannan (1988), above Theorem 7.4.5, and is equal to

$$
\dot{\sigma}_n^2 \;=\; \frac{1}{n} \sum_{t=1}^{n} \varepsilon_t^2 \;,\; n \in I\!N \;. \qquad (5.4)
$$

Summarizing we obtain from Theorem 5.1 the following expansion, which holds uniformly in $p \in \{1, ..., p(n)\}$

$$
\log \left\{ \tilde{\sigma}_n^2(p) + 2 \cdot S_n^*(p \right\} = \log \dot{\sigma}_n^2 + \left\{ \frac{p}{n} + \frac{\sigma^2(p) - \sigma^2}{\sigma^2} \right\} (1 + o_P(1)) \;.
$$

This is exactly the same expansion as Deistler and Hannan obtained for the AIC, cf. Deistler and Hannan (1988), Theorem 7.4.7. In other words, the considerations given below Theorem 7.4.7 in Deistler and Hannan (1988) hold also true for the bootstrap order selection.

Remark . (i) Shibata has a result similar to Theorem 5.1 in his paper, cf. Shibata (1980) Lemma 7.1, but we can dispend with the assumption of normality.

(ii) From Theorem 5.1 we obtain exactly along the lines of Section 4 in Shibata (1980) the *asymptotic efficiency* of the bootstrap order selection under the same assumptions as in Shibata. The concept of asymptotic efficiency is also defined by Shibata.
Following the arguments given in Deistler and Hannan (1988), p. 333/334, we obtain exactly along the same lines and under the same assumption that

$$\frac{P_B(n)}{\operatorname{argmin}_{1\le p\le p(n)}\left(\frac{p}{n}+\frac{\sigma^2(p)-\sigma^2}{\sigma^2}\right)} \to 1 \quad \text{in probability.}$$

In the next Section we report some simulation results for the bootstrap order selection in comparison with other order selection procedures.

6. Simulations

Let us consider the following two order selection procedures for a simulation study. The argument of the minimum (argmin) is in both cases computed over the range $\{1, ..., p(n)\}$.

$$\mathbf{AIC} \;=\; \operatorname{argmin}_p \left\{\hat{\sigma}_n^2(p)\cdot\left(1+\frac{2p}{n}\right)\right\} \tag{6.1}$$

$$\mathbf{P_B} \;=\; \operatorname{argmin}_p \left\{\hat{\sigma}_n^2(p) + 2\cdot E^*\left\|\hat{a}^*(p)-\tilde{a}(p)\right\|^2_{\tilde{R}(p)}\right\} . \tag{6.2}$$

In all cases

$$\hat{\sigma}_n^2(p) = 1 - \sum_{\ell=1}^{p}\hat{a}_\ell\hat{r}_\ell$$

and $\hat{r}_h$ denotes an estimator of the autocorrelation at lag h, which does not necessarily have to coincide with the empirical autocorrelation $\tilde{r}_h$ of the observations. This deviates slightly from the preceding sections, where we preferred to work with the autocovariances in order to simplify the proofs, and obviates the need for an M-estimator of γ_0. The AIC goes back to Akaike (1973a,b, 1974). P_B denotes the bootstrap order selection proposed in Section 4 of the present paper. Note that for the theoretical investigation we used a slightly modified version of P_B.

criterion	p=1	p=2	p=3	p=4	p=5	p=6	p=7	p=8
AIC	2	0	67	14	8	3	3	3
AIC	3	0	72	16	4	4	1	0
AIC	4	0	69	13	5	6	2	1
$\mathbf{P_B}, \psi_{\text{id}}$	1	1	68	13	6	4	2	5
$\mathbf{P_B}, \psi_{\text{id}}$	5	1	70	12	5	3	2	2
$\mathbf{P_B}, \psi_{\text{id}}$	5	0	63	14	9	3	3	3
$\mathbf{P_B}, \psi_{\text{Huber}}$	1	1	68	10	7	8	3	2
$\mathbf{P_B}, \psi_{\text{Huber}}$	3	0	66	17	3	6	2	3
$\mathbf{P_B}, \psi_{\text{Huber}}$	3	0	65	12	9	4	3	4

Table 6.1
frequencies of selected orders (100 repetitions)
model (6.3), normal innovations, sample size=100, p(n)=8

The simulations we are going to report are based on the following three stationary time series models

$$\begin{aligned} X_t &= 0.64 \cdot X_{t-1} - 0.19 \cdot X_{t-2} + 0.39 \cdot X_{t-3} + \varepsilon_t & (6.3)\\ X_t &= -X_{t-2} - 0.1 \cdot X_{t-4} + \varepsilon_t & (6.4)\\ X_t &= -0.5 \cdot X_{t-2} + 0.5 \cdot \varepsilon_{t-1} + \varepsilon_t \ . & (6.5)\end{aligned}$$

The first two models are of finite autoregressive order, while the ARMA(2,1)-model (6.5) possesses an autoregressive representation of infinite order.

For the innovations ε_t we use the following distributions

$$\begin{aligned} \varepsilon_1 &\sim \mathcal{N}(0,1) \quad \text{normally distributed innovations} & (6.6)\\ \varepsilon_1 &\sim 0.8\mathcal{N}(0,1) + 0.2\mathcal{N}(0,25) \qquad \text{contaminated innovations} & (6.7)\\ \varepsilon_1 &\sim 0.5\left(\mathcal{N}(-3,1) + \mathcal{N}(3,1)\right) \qquad \text{bimodal normal innovations} & (6.8)\end{aligned}$$

The AIC is always computed using least squares parameter estimates, for which this criterion is designed. However, changing the parameter estimates does not affect the AIC essentially. The bootstrap order selection P_B is computed for different M-estimators. Here we make use of $\psi_{\text{id}}(x) = x$, corresponding to least squares, and $\psi_{\text{Huber}}(x) = -\mathbf{1}_{(x<-1)} + x \cdot \mathbf{1}_{(-1 \le x \le 1)} + \mathbf{1}_{(x>1)}$.
We report on the simulated behaviour of the procedures on two different random samples of 100 time series each in order to give an impression of the stochastic fluctuation of the results.

criterion	p=1	p=2	p=3	p=4	p=5	p=6	p=7	p=8
AIC	0	71	7	12	4	2	0	4
AIC	0	61	11	15	7	2	2	2
AIC	0	65	13	13	3	4	1	1
$\mathbf{P_B}, \psi_{\text{id}}$	0	68	5	16	5	2	2	2
$\mathbf{P_B}, \psi_{\text{id}}$	0	59	10	13	6	3	3	6
$\mathbf{P_B}, \psi_{\text{id}}$	0	70	6	17	2	4	0	1
$\mathbf{P_B}, \psi_{\text{opt}}$	0	55	3	37	4	0	1	0
$\mathbf{P_B}, \psi_{\text{opt}}$	0	67	0	30	1	1	0	1
$\mathbf{P_B}, \psi_{\text{opt}}$	0	62	6	27	5	0	0	0

Table 6.2
frequencies of selected orders (100 repetitions)
model (6.4), bimodal innovations, sample size=100, p(n)=8

Tables 6.1 gives the results for model (6.3). From this table it can be seen that the results for normally distributed observations do not differ very much. This means that the proposed bootstrap order selection procedure behaves more or less like the AIC for standard situations. For non-normally distributed innovations the situation is quite different. To demonstrate this let us first consider model (6.4) with bimodal normally distributed innovations and sample sizes $n = 100$ (cf. Table 6.2) and $n = 200$ (cf. Table 6.3). Here we make use of the asymptotically optimal choice of the ψ-function, namely ψ equal to the logarithmic derivative of the underlying density, i.e. $\psi = -f'/f$. Additionally we present results for the least-squares estimator, i.e. $\psi(x) = x$.

criterion	p=1	p=2	p=3	p=4	p=5	p=6	p=7	p=8
AIC	0	55	8	26	4	2	0	5
AIC	0	57	2	32	3	2	0	4
AIC	0	45	5	31	6	7	4	2
$\mathbf{P_B}, \psi_{\text{id}}$	0	51	8	26	4	6	4	1
$\mathbf{P_B}, \psi_{\text{id}}$	0	52	7	24	10	4	2	1
$\mathbf{P_B}, \psi_{\text{id}}$	0	59	3	28	7	0	1	2
$\mathbf{P_B}, \psi_{\text{opt}}$	0	9	0	79	7	2	3	0
$\mathbf{P_B}, \psi_{\text{opt}}$	0	14	0	74	9	2	0	1
$\mathbf{P_B}, \psi_{\text{opt}}$	0	5	0	81	9	3	0	2

Table 6.3
frequencies of selected orders (100 repetitions)
model (6.4), bimodal innovations, sample size=200, p(n)=8

criterion	p=1	p=2	p=3	p=4	p=5	p=6	p=7	p=8	p=9	p=10
AIC	0	1	22	44	18	11	0	0	1	3
AIC	0	1	17	42	20	10	5	3	1	1
AIC	0	0	18	52	17	3	4	5	0	1
$\mathbf{P_B}, \psi_{\text{Huber}}$	0	0	4	31	29	21	8	2	4	1
$\mathbf{P_B}, \psi_{\text{Huber}}$	0	0	3	30	29	18	7	5	5	3
$\mathbf{P_B}, \psi_{\text{Huber}}$	0	0	4	41	30	11	8	2	4	0
$\mathbf{P_B}, \psi_{\text{opt}}$	0	0	2	40	38	6	8	3	2	1
$\mathbf{P_B}, \psi_{\text{opt}}$	0	0	1	39	40	13	2	2	0	3
$\mathbf{P_B}, \psi_{\text{opt}}$	0	0	3	37	29	11	9	3	8	0

Table 6.4
frequencies of selected orders (100 repetitions)
model (6.5), contaminated innovations, sample size=200, p(n)=200

It can be seen clearly, especially from Table 6.3, that the bootstrap order selection using the asymptotically optimal ψ-function tends to select the true order with much higher probability. This is due to the fact that M-estimators with this ψ-function have much smaller variance than, for example, the least squares estimator used in the construction of the AIC.

Finally, for the ARMA(2,1)-model (6.5) we again demonstrate the behaviour of the bootstrap order selection for two different M-estimators (ψ_{Huber} and $\psi_{\text{opt}} = -f'/f$) and contaminated innovations (cf. Table 6.4). The precision of the parameter estimates increases from the Huber M-estimator to M-estimates with asymptotically optimal score-function, which implies the desired property that the P_B tends to higher orders for the autoregressive fit.

Acknowledgement. Parts of the research presented in this paper was done while the second author enjoyed the hospitality of the Sonderforschungsbereich 123 at the University of Heidelberg which is gratefully acknowledged. The third author acknowledges the support of the DFG-project MA 1026/6-1.

7. Proofs

The proof of Theorem 4.1 will be based on the following approximation lemma, which is of interest on its own.

Lemma 7.1 : *Under the assumptions of Theorem 4.1 we have*

$$\max_{1\le h,k\le n} |\mathrm{Cov}^* (\tilde{\gamma}_h^*, \tilde{\gamma}_k^*) - \mathrm{Cov}\,(\tilde{\gamma}_h, \tilde{\gamma}_k)| \quad = \quad o_P(n^{-1}), \tag{7.1}$$

$$\max_{1\le h,k\le n} |\mathrm{Cov}\,(\tilde{\gamma}_h, \tilde{\gamma}_k)| \quad = \quad O(n^{-1}). \tag{7.2}$$

Proof : We will show the following inequality

$$\max_{1\le h,k\le n} \left|\mathrm{Cov}^*\left(\tilde\gamma_h^*, \tilde\gamma_k^*\right) - \mathrm{Cov}\left(\tilde\gamma_h, \tilde\gamma_k\right)\right| \;\le\; \frac{3}{n}\left\|\boldsymbol\alpha - \boldsymbol\alpha^*\right\|_1\left(\|\boldsymbol\alpha\|_1 + \|\boldsymbol\alpha^*\|_1\right)^3 E\,\varepsilon_0^4$$
$$+ \frac{3}{n}\|\boldsymbol\alpha^*\|_1^4\left(\left|\left(E\,\varepsilon_0^2\right)^2 - \left(E^*\,\varepsilon_0^{*2}\right)^2\right| + \left|E\,\varepsilon_0^4 - E^*\,\varepsilon_0^{*4}\right|\right) \tag{7.3}$$

where $\boldsymbol\alpha^* = \tilde{\boldsymbol\alpha}(p_0) = \mathbf{j}(\tilde a(p_0))^{-1}$. The asymptotic properties of the bootstrap construction mentioned in Section 4 will then imply (7.1), and (7.2) will follow from (7.3) by setting $\boldsymbol\alpha^* = \boldsymbol\gamma^* = \mathbf{O}$.

The proof of (7.3) will be based on the MA(∞)-representation of the process (X_t), which yields the following formula for the empirical autocovariances:

$$\begin{aligned}\tilde\gamma_h - E\,\tilde\gamma_h &= \frac{1}{n}\sum_{t=1}^{n-h}\sum_{j,\ell=0}^{\infty}{}^{\dagger}\,\alpha_j\alpha_\ell\varepsilon_{t-j}\varepsilon_{t+h-\ell} + \frac{1}{n}\sum_{t=1}^{n-h}\sum_{j=0}^{\infty}\alpha_j\alpha_{j+h}\left(\varepsilon_{t-j}^2 - E\,\varepsilon_0^2\right)\\ &=: V(h) + W(h),\end{aligned}$$

where the dagger indicates that summation takes place only over those pairs (j,ℓ) with $\ell \ne j+h$. For any h_1, h_2, the sums $V(h_1)$ and $W(h_2)$ are uncorrelated which implies $\mathrm{Cov}\left(\tilde\gamma_{h_1}, \tilde\gamma_{h_2}\right) = E\,V(h_1)V(h_2) + E\,W(h_1)W(h_2)$. Furthermore

$$\begin{aligned}E\,V(h_1)V(h_2) &= \frac{1}{n^2}E\prod_{i=1}^{2}\left(\sum_{t=1}^{n-h_i}\sum_{j,\ell=0}^{\infty}{}^{\dagger}\,\alpha_j\alpha_\ell\varepsilon_{t-j}\varepsilon_{t+h_i-\ell}\right)\\ &= \frac{1}{n^2}\sum_{j_1,\ell_1=0}^{\infty}{}^{\dagger}\sum_{j_2,\ell_2=0}^{\infty}{}^{\dagger}\prod_{i=1}^{2}\left(\alpha_{j_i}\alpha_{\ell_i}\right)\sum_{t_1=1}^{n-h_1}\sum_{t_2=1}^{n-h_2}E\prod_{i=1}^{2}\left(\varepsilon_{t_i-j_i}\varepsilon_{t_i+h_i-\ell_i}\right).\end{aligned}$$

As $j_i \ne h_i+\ell_i$, the last expectation equals zero unless $t_1-j_1 = t_2-j_2$ or $t_1-j_1 = t_2+h_2-\ell_2$, in which case it may be $(E\varepsilon_0^2)^2$ instead of zero. Hence the double sum over t_1, t_2 reduces to two single sums over t_1. Taking the difference $EV(h_1)V(h_2) - E^*V^*(h_1)V^*(h_2)$ we first replace $\boldsymbol\alpha$ by $\boldsymbol\alpha^*$ to get

$$\begin{aligned}&\left|\frac{1}{n^2}\sum_{j_1,\ell_1=0}^{\infty}{}^{\dagger}\sum_{j_2,\ell_2=0}^{\infty}{}^{\dagger}\left(\prod_{i=1}^{2}\alpha_{j_i}\alpha_{\ell_i} - \prod_{i=1}^{2}\alpha_{j_i}^*\alpha_{\ell_i}^*\right)\sum_{t_1=1}^{n-h_1}\sum_{t_2=1}^{n-h_2}E\prod_{i=1}^{2}\left(\varepsilon_{t_i-j_i}\varepsilon_{t_i+h_i-\ell_i}\right)\right|\\ \le\;& \frac{2}{n}\left\|\boldsymbol\alpha - \boldsymbol\alpha^*\right\|_1\left(\|\boldsymbol\alpha\|_1 + \|\boldsymbol\alpha^*\|_1\right)^3\left(E\,\varepsilon_0^2\right)^2.\end{aligned} \tag{7.4}$$

In a second step the innovations ε_t are exchanged for the bootstrap innovations ε_t^*, yielding

$$\begin{aligned}&\left|\frac{1}{n^2}\sum_{j_1,\ell_1=0}^{\infty}{}^{\dagger}\sum_{j_2,\ell_2=0}^{\infty}{}^{\dagger}\prod_{i=1}^{2}\alpha_{j_i}^*\alpha_{\ell_i}^*\sum_{t_1=1}^{n-h_1}\sum_{t_2=1}^{n-h_2}E\prod_{i=1}^{2}\varepsilon_{t_i-j_i}\varepsilon_{t_i+h_i-\ell_i} - E^*\prod_{i=1}^{2}\varepsilon_{t_i-j_i}^*\varepsilon_{t_i+h_i-\ell_i}^*\right|\\ \le\;& \frac{2}{n}\|\boldsymbol\alpha^*\|_1^4\left|\left(E\,\varepsilon_0^2\right)^2 - \left(E\,\varepsilon_0^{*2}\right)^2\right|.\end{aligned} \tag{7.5}$$

A bound for $E\,V(h_1)V(h_2) - E^*\,V^*(h_1)V^*(h_2)$ is obtained by adding the bounds in (7.4) and (7.5). A similar calculation leads to

$$|E\,W(h_1)W(h_2) - E^*\,W^*(h_1)W^*(h_2)| \le \frac{1}{n}\,\|\boldsymbol{\alpha} - \boldsymbol{\alpha}^*\|_1\left(\|\boldsymbol{\alpha}\|_1 + \|\boldsymbol{\alpha}^*\|_1\right)^3 E\,\varepsilon_0^4$$
$$+\ \frac{1}{n}\,\|\boldsymbol{\alpha}^*\|_1^4\left|E\left(\varepsilon_0^2 - E\,\varepsilon_0^2\right)^2 - E^*\left(\varepsilon_0^{*2} - E^*\,\varepsilon_0^{*2}\right)^2\right|.$$

This proves the lemma.

Proof of Theorem 4.1 : We first note that $S_n(p)$ and $S_n^*(p)$, cf. (2.8) and (4.1), can be written as

$$S_n(p) \;=\; E\,\|A(p)\,(\tilde{\gamma}_1(p) - \gamma_1(p))\|^2_{\Gamma(p)^{-1}}$$

and

$$S_n^*(p) \;=\; E^*\,\|A^*(p)\,(\tilde{\gamma}_1^*(p) - \gamma_1^*(p))\|^2_{\Gamma^*(p)^{-1}}\ ,$$

where the $p \times 2p$ -matrix $A(p)$ is defined as

$$\begin{bmatrix} 0 & 0 & \cdots & 0 & 0 & 1 & -a_1(p) & \cdots & -a_{p-1}(p) & -a_p(p) \\ 0 & 0 & \cdots & 0 & 1 & -a_1(p) & -a_2(p) & \cdots & -a_p(p) & 0 \\ \vdots & & & \ddots & & & & \ddots & & \vdots \\ 0 & 1 & -a_1(p) & \cdots & & & -a_p(p) & 0 & \cdots & 0 \\ 1 & -a_1(p) & -a_2(p) & \ldots & & -a_p(p) & 0 & \cdots & & 0 \end{bmatrix},$$

$\gamma_1(p) := (\gamma_p, \gamma_{p-1}, \ldots, \gamma_{1-p})^T \in \mathbb{R}^{2p}$ with $\gamma_{-h} = \gamma_h$, $A^*(p)$ is defined as $A(p)$ with $\tilde{a}_\nu(p)$ instead of $a_\nu(p)$ and $\tilde{\gamma}_1(p)$, $\tilde{\gamma}_1^*(p)$, $\gamma_1^*(p)$ are defined analogously to $\gamma_1(p)$.

Writing $\Sigma(p) = A(p)^T\Gamma(p)^{-1}A(p)$, $\Sigma^*(p) = A^*(p)^T\Gamma^*(p)^{-1}A^*(p)$, we have

$$\begin{aligned} |S_n^*(p) \;-\; S_n(p)| \;&=\; \left|E^*\,\|\tilde{\gamma}_1^*(p) - \gamma_1^*(p)\|^2_{\Sigma^*(p)} \;-\; E\,\|\tilde{\gamma}_1(p) - \gamma_1(p)\|^2_{\Sigma(p)}\right| \\ &\le\; \left|E^*\,\|\tilde{\gamma}_1^*(p) - \gamma_1^*(p)\|^2_{\Sigma^*(p)} \;-\; E^*\,\|\tilde{\gamma}_1^*(p) - \gamma_1^*(p)\|^2_{\Sigma(p)}\right| \\ &\quad +\; \left|E^*\,\|\tilde{\gamma}_1^*(p) - \gamma_1^*(p)\|^2_{\Sigma(p)} \;-\; E\,\|\tilde{\gamma}_1(p) - \gamma_1(p)\|^2_{\Sigma(p)}\right|\ . \end{aligned}$$

Bound the first summand through $\|\Sigma^*(p) \;-\; \Sigma(p)\| \cdot E^*\,\|\tilde{\gamma}_1^*(p) - \gamma_1^*(p)\|_2^2$. Because of Lemma 7.1 it suffices to show that

$$\begin{aligned} \max_p \|\Sigma^*(p) \;-\; \Sigma(p)\| \;&=\; \max_p \left\|A^*(p)^T\Gamma^*(p)^{-1}A^*(p) \;-\; A(p)^T\Gamma(p)^{-1}A(p)\right\| \\ &=\; o_P(1)\,. \end{aligned}$$

As $\|\Gamma(p)\|$ and $\|\Gamma(p)^{-1}\|$ are uniformly bounded in $p \in \mathbb{N}$, the last equation will follow from

$$\max_p \|A^*(p) - A(p)\|\,,\ \max_p \left\|\Gamma^*(p)^{-1} - \Gamma(p)^{-1}\right\| = o_P(1), \max_p \|A(p)\| = O(1).$$

In view of the matrix inequalities $\|B\| \leq \|B\|_2 := (\sum |b_{ij}|^2)^{\frac{1}{2}}$ and $\|B^{-1} - C^{-1}\| \leq \frac{\|B^{-1}\|^2\|B-C\|}{1-\|B^{-1}\|\|B-C\|}$ if $\|B^{-1}\|\,\|B-C\| < 1$, we obtain from Lemma 7.1

$$\max_p \left\|\tilde{\Gamma}(p) - \Gamma(p)\right\| = o_P(1)\,, \tag{7.6}$$

$$\max_p \left\|\tilde{\Gamma}(p)^{-1} - \Gamma(p)^{-1}\right\| = o_P(1)\,. \tag{7.7}$$

As $\Gamma^*(p) = \frac{\gamma_0^*}{\tilde{\gamma}_0}\tilde{\Gamma}(p)$ and $\frac{\gamma_0^*}{\tilde{\gamma}_0} \to 1$ in probability, (7.6) and (7.7) hold with $\tilde{\Gamma}(p)$ replaced by $\Gamma(p)^*$.

Now, for any $x \in I\!R^{2p}$ the vector $A(p)x$ obviously consists of certain entries of the convolution $\mathbf{j}(a(p)) * x^-$ where $\mathbf{j}(a(p)), x$ are embedded in $I\!R^{\mathbb{Z}}$, $(x^-)_h = x_{-h}$ and convolution takes place over $\mathbb{Z}$. From the convolution inequality $\|\mathbf{b} * \mathbf{c}\|_2 \leq \|\mathbf{b}\|_2\,\|\mathbf{c}\|_1$ we conclude

$$\|A(p)\| = \sup_{\|x\|_2=1} \|A(p)x\|_2 \;\leq\; 1 + \|a(p)\|_1\,, \tag{7.8}$$

where the latter is bounded uniformly in $p \in I\!N$ according to Theorem 2.2 of Baxter (1962). As by Lemma 7.1 and (7.7)

$$\begin{aligned} \max_{1\leq p\leq p(n)} \|\tilde{a}(p) - a(p)\|_2 &= \max_p \left\|\tilde{\Gamma}(p)^{-1}\tilde{\gamma}(p) - \Gamma(p)\gamma(p)\right\|_2 \\ &= O_P\left(\sqrt{\frac{p(n)^3}{n}}\right) = o_P\left(\frac{1}{\sqrt{p(n)}}\right)\,, \end{aligned} \tag{7.9}$$

we get in the same manner as in (7.8)

$$\max_p \|A^*(p) - A(p)\| \leq \max_p \|\tilde{a}(p) - a(p)\|_1 = o_P(1)\,.$$

For the second summand we have

$$\begin{aligned} &\left|E^*\,\|\tilde{\gamma}_1^*(p) - \gamma_1^*(p)\|^2_{\Sigma(p)} \;-\; E\,\|\tilde{\gamma}_1(p) - \gamma_1(p)\|^2_{\Sigma(p)}\right| \\ \leq\;& \sum_{h,k=1}^{p} |\Sigma(p)_{hk}| \left|\mathrm{Cov}^*\left(\tilde{\gamma}_h^*, \tilde{\gamma}_k^*\right) - \mathrm{Cov}\left(\tilde{\gamma}_h, \tilde{\gamma}_k\right) + \frac{hk}{n^2}\left(\tilde{\gamma}_h\tilde{\gamma}_k - \gamma_h\gamma_k\right)\right|\,. \end{aligned}$$

Now, from Shibata (1980), p. 151, there is a constant C such that $\sum_{h,k=1}^p |\Gamma(p)^{-1}_{h,k}| \leq Cp$ for all $p \in I\!N$. Hence we have $\sum_{h,k=1}^p |\Sigma(p)_{hk}| \leq Cp(1 + \|a(p)\|_1)^2$, and by Lemma 7.1 and the uniform boundedness of $\|a(p)\|_1$,

$$\max_p \frac{n}{p}\left|E^*\,\|\tilde{\gamma}_1^*(p) - \gamma_1^*(p)\|^2_{\Sigma(p)} \;-\; E\,\|\tilde{\gamma}_1(p) - \gamma_1(p)\|^2_{\Sigma(p)}\right| \;=\; o_P(1).$$

This concludes the proof of Theorem 4.1.

Proof of Theorem 5.1 : Because of Theorem 4.1 it suffices to consider

$$\max_{1\le p\le p(n)} \frac{\left|S_n(p) - \frac{p}{n}\sigma^2\right|}{\frac{p}{n} + \frac{\sigma^2(p)-\sigma^2}{\sigma^2}}\,. \tag{7.10}$$

Using convention (5.1), (5.2) we obtain

$$S_n(p) \;=\; E\left\|\tilde\gamma(p) - \tilde\Gamma(p)a(p)\right\|^2_{\Gamma(p)^{-1}} \;=\; E\left\|\frac{1}{n}\sum_t \varepsilon_t(a(p))X_{t-1}(p)\right\|^2_{\Gamma(p)^{-1}}$$

where $\varepsilon_t(a(p)) = X_t - a(p)^T X_{t-1}(p)$. Now, since

$$E\left\|\frac{1}{n}\sum_t \varepsilon_t X_{t-1}(p)\right\|^2_{\Gamma(p)^{-1}} \;=\; \frac{1}{n^2}\sum_t E\varepsilon_t^2\; E\left(X_{t-1}(p)^T\Gamma(p)^{-1}X_{t-1}(p)\right) \tag{7.11}$$

$$=\; \frac{n-p(n)}{n}\frac{p}{n}\sigma^2\,, \tag{7.12}$$

we have to show

$$\max_p \frac{\left|E\left\|\frac{1}{n}\sum_t \varepsilon_t(a(p))X_{t-1}(p)\right\|^2_{\Gamma(p)^{-1}} - E\left\|\frac{1}{n}\sum_t \varepsilon_t X_{t-1}(p)\right\|^2_{\Gamma(p)^{-1}}\right|}{\frac{p}{n} + \frac{\sigma^2(p)-\sigma^2}{\sigma^2}} \;=\; o_P(1)\,.$$

To this end consider

$$\begin{aligned}
&\left|E\left\|\frac{1}{n}\sum_t \varepsilon_t(a(p))X_{t-1}(p)\right\|^2_{\Gamma(p)^{-1}} - E\left\|\frac{1}{n}\sum_t \varepsilon_t X_{t-1}(p)\right\|^2_{\Gamma(p)^{-1}}\right| \\
&\le\; E\left\|\frac{1}{n}\sum_t \left(\varepsilon_t(a(p)) - \varepsilon_t\right)X_{t-1}(p)\right\|^2_{\Gamma(p)^{-1}} \\
&\quad +2\sqrt{E\left\|\frac{1}{n}\sum_t \varepsilon_t X_{t-1}(p)\right\|^2_{\Gamma(p)^{-1}}}\sqrt{E\left\|\frac{1}{n}\sum_t \left(\varepsilon_t(a(p)) - \varepsilon_t\right)X_{t-1}(p)\right\|^2_{\Gamma(p)^{-1}}}\,.
\end{aligned}$$

Because of (7.12), we may restrict our attention to

$$\max_p \frac{E\left\|\frac{1}{n}\sum_t \left(\varepsilon_t(a(p)) - \varepsilon_t\right)X_{t-1}(p)\right\|^2_{\Gamma(p)^{-1}}}{\frac{p}{n} + \frac{\sigma^2(p)-\sigma^2}{\sigma^2}}\,.$$

Now, $\|\Gamma(p)\|$ and $\|\Gamma(p)^{-1}\|$ are uniformly bounded in $p \in I\!N$, $\sigma^2(p) - \sigma^2 = \|\mathbf{a} - a(p)\|^2_\Gamma$ and

$$\sqrt{E\left\|\frac{1}{n}\sum_t \left(\varepsilon_t(a(p)) - \varepsilon_t\right)X_{t-1}(p)\right\|^2_2} \le \sum_{h=1}^{p}\sqrt{E\left[\frac{1}{n}\sum_t \left(\varepsilon_t(a(p)) - \varepsilon_t\right)X_{t-h}\right]^2}\,.$$

Therefore it suffices to show

$$E\left[\frac{1}{n}\sum_t \left(\varepsilon_t(a(p)) - \varepsilon_t\right)X_{t-h}\right]^2 \;\le\; \frac{C}{n}\left\|\mathbf{a} - a(p)\right\|_2^2, \qquad h = 1,\ldots,p, \tag{7.13}$$

where $C > 0$ is a constant independent of p. For this purpose, we fix p and write $Z_t := \varepsilon_t(a(p)) - \varepsilon_t$. Then $Z_t = \sum_{\ell=0}^{\infty} \beta_\ell \varepsilon_{t-\ell}$ where $\boldsymbol{\beta} = (\mathbf{j}(a(p)) - \mathbf{j}(\mathbf{a})) * \boldsymbol{\alpha}$ and $\boldsymbol{\alpha} = \mathbf{a}^{-1}$. With these notations, (7.13) becomes

$$\frac{1}{n^2} \sum_{s,t} E\ Z_s X_{s-h} Z_t X_{t-h} \ \leq\ \frac{C}{n} \|\mathbf{a} - a(p)\|_2^2, \quad h = 1, \ldots, p. \tag{7.14}$$

Using the MA(∞)–representations of the processes $(X_t), (Z_t)$ and the orthogonality $E\, Z_t X_{t-h} = 0$, $h = 1, \ldots, p$, it is easy to see that for $s \leq t$

$$\begin{aligned} E\, Z_s X_{s-h} Z_t X_{t-h} \ &=\ E\, Z_s Z_t\, E\, X_{s-h} X_{t-h} + E\, Z_s X_{t-h}\, E\, Z_t X_{s-h} \\ &\quad + \sum_{\ell=0}^{\infty} \beta_{\ell+h} \alpha_\ell \beta_{\ell+t-s+h} \alpha_{\ell+t-s} \left(E\, \varepsilon_0^4 - 3\sigma^4\right) . \end{aligned} \tag{7.15}$$

Setting $\alpha_\ell, \beta_\ell = 0$ if $\ell < 0$ we note that for any $k \in \mathbb{Z}$

$$E\, Z_0 X_k = \sum_{\ell=-\infty}^{\infty} \beta_\ell \alpha_{\ell+k} = \sum_{\ell=-\infty}^{\infty} \beta_\ell \alpha_{-(-\ell-k)} = \left(\boldsymbol{\beta} * \boldsymbol{\alpha}^-\right)_{-k}$$

where $\boldsymbol{\alpha}^- = (\alpha_\ell : \ell \in \mathbb{Z})$ and convolution takes place over $\mathbb{Z}$. Similar expressions may be derived for the autocovariances $E\, Z_0 Z_k$ and $E\, X_0 X_k$. Summation of the first term on the right side of (7.15) over $t \geq s$ gives

$$\begin{aligned} \sum_{t \geq s} |E\, Z_s Z_t\, E\, X_{s-h} X_{t-h}| \ &=\ \sum_{t \geq 0} \left| \left(\boldsymbol{\beta} * \boldsymbol{\beta}^-\right)_t \left(\boldsymbol{\alpha} * \boldsymbol{\alpha}^-\right)_t \right| \\ &\leq\ \left\|\boldsymbol{\beta} * \boldsymbol{\beta}^-\right\|_2 \left\|\boldsymbol{\alpha} * \boldsymbol{\alpha}^-\right\|_2 \\ &\leq\ \|\mathbf{a} - a(p)\|_2^2 \, \|\boldsymbol{\alpha}\|_1^4 \end{aligned}$$

by repeated use of the convolution inequality $\|\mathbf{b} * \mathbf{c}\|_2 \leq \|\mathbf{b}\|_2 \|\mathbf{c}\|_1$ and $\|\mathbf{b}^-\|_1 = \|\mathbf{b}\|_1$. Summation of the other two terms on the right side of (7.15) leads to similar expressions, so we may conclude (7.14).

References

Akaike, H. (1973b). Information theory and an extension of the maximum likelihood principle. *Second International Symposium on Information Theory (B.N. Petrov and F. Csaki, eds.)*, Akademia Kiado, Budapest, 267-281.

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

Baxter, G. (1962). An asymptotic result for the finite predictor. *Math. Scand.* **10**, 137-144.

Behrens, J. (1990). Robust order selection for autoregressive processes. *Dissertation.* Universität Kaiserslautern.

Brockwell, P.J. and Davis, R.A. (1991). Time Series: Theory and Methods (Second

Edition). Springer-Verlag, New York.
Bühlmann, P. (1997). Sieve bootstrap for time series. *Bernoulli* **3**, 123-148.
Deistler, M. and Hannan, E.J. (1988). The Statistical Theory of Linear Systems. John Wiley, New York.
Kreiss, J.-P. (1988). Asymptotic statistical inference for a class of stochastic processes. *Habilitationsschrift.* Univ. of Hamburg.
Kreiss, J.-P. (1997). Asymptotical properties of residual bootstrap for autoregressions. *J. Time Ser. Anal.* (submitted).
Martin, R.D. (1983). The Cramér–Rao bound and robust M–estimates for autoregressions. *Biometrika* **69**, 437-442.
Moser, M. (1997). Bootstrap–Ordnungswahl und M–Schätzung in linearer Autoregression. *Dissertation.* TU Braunschweig.
Pedersen, G.K. (1995). Analysis Now (Second Edition). Springer-Verlag, New York.
Shibata, R. (1980). Asymptotically efficient selection of the order of the the model for estimating parameters of a linear process. *Ann. Statist.* **8**, 147-164.

Jürgen Franke
Fachbereich Mathematik
Universität Kaiserslautern
Erwin-Schrödinger-Strasse
67663 Kaiserslautern
Germany

Jens-Peter Kreiss
Institut für Mathematische Stochastik
Technische Universität Braunschweig
Pockelstrasse 14
38106 Braunschweig
Germany

Martin Moser
Institut für Angewandte Mathematik
Universität Heidelberg
Im Neuenheimer Feld 294
69120 Heidelberg
Germany