Econometric Theory, 7, 1991, 163–185. Printed in the United States of America.

EFFECTS OF MODEL SELECTION ON INFERENCE

B.M. PÖTSCHER
University of Maryland

The asymptotic properties of parameter estimators which are based on a model that has been selected by a model selection procedure are investigated. In particular, the asymptotic distribution is derived and the effects of the model selection process on subsequent inference are illustrated.

1. INTRODUCTION

When fitting a model to data, the choice of the model itself is often based on the same data set. For example, in a regression context test procedures or model selection criteria like Mallows' [20] C_p, Akaike's [2] AIC, or Schwarz's [28] BIC may be used to select an appropriate model, namely, set of regressors. Similarly, the number of lags in an autoregressive model or in a distributed lag model is frequently determined by means of such procedures. The statistical properties of some of these procedures have been analyzed in the recent statistics literature, especially in the context of order estimation of autoregressive moving average models and of regressor selection. This research has been concerned primarily with the construction of model selection procedures and with optimality results for the selected model [1,2,3,5,7,28,31], or with consistency of the model selection procedure, that is, with the question of whether a "correct" and "minimal" model is chosen asymptotically [6,8,11,12,13,14,21,22,23,25,27,35]. The important question of how the use of a model selection procedure affects the asymptotic distribution of parameter estimators and related statistics, and hence subsequent inference, does not seem to have been studied in the literature to the same extent. The goal of the paper is to shed some light on this problem and to derive the relevant asymptotic distributions.[1] In Section 2 some simple preliminary results are collected covering in particular the case of consistent model selection procedures. Section 3 presents the asymptotic distribution of M-estimators (including quasi-maximum likelihood estimators), when the model is selected by means of a multiple testing procedure. In this section also a numerical example is presented to illustrated the theoretical results. Section 4 contains remarks on model selection, in particular on the relation

Helpful comments were provided by P.M. Robinson, the referees, and by participants in the Econometrics Seminar at Yale University. Financial support by the Österreichischen Fonds zur Förderung der Wissenschaftlichen Forschung, Grant No. S32/02, is gratefully acknowledged. Thanks are also due to C. Weber for his excellent computational assistance.

0266-4666/91 $5.00 + .00

between multiple testing procedures and model selection procedures based on criteria like AIC, as well as on the relative merits of consistent and inconsistent procedures. All proofs are given in the appendix.

2. PRELIMINARIES

Let $(M(p):p \in \mathcal{P})$ be a family of models for a data-generating process where the index set $\mathcal{P}$ is a nonvoid, finite, or countable set, $\varnothing \neq \mathcal{P} \subseteq \mathbb{N}_0$ without loss of generality where $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$. In this section we do not assume that the models are necessarily nested. Each model $M(p)$ will typically be specified only up to some unknown parameter of finite or infinite dimension. For example, in the context of autoregressive models $M(p)$ would stand for the autoregressive model of order p. A model selection procedure is a rule associating with any sample of size n an element $\hat{p}_n$ of $\mathcal{P}$. Given $\hat{p}_n$, inference is then usually based on the model $M(\hat{p}_n)$. Typically, given the data-generating process, there is an element $p_0 \in \mathcal{P}$ corresponding to a unique "minimal" and "true" model; hence, $\hat{p}_n$ can be regarded as an estimator for p_0. For example, in order estimation in autoregressive models $\mathcal{P}$ is $\mathbb{N}_0$, or an interval thereof, and p_0 is the order of the data-generating autoregressive process, that is, the index of the highest lag with a nonzero coefficient in the true structure. The following lemma establishes the rather obvious fact that inference based on the selected model $M(\hat{p}_n)$ is asymptotically identical with inference based on $M(p_0)$ if $\hat{p}_n$ is consistent for p_0. In the following, $T(n,p)$ denotes any statistic based on model $M(p)$.

LEMMA 1. *If $\hat{p}_n$ is consistent for p_0, that is,* $\mathrm{pr}(\hat{p}_n = p_0) \rightarrow 1$ *as $n \rightarrow \infty$, then the statistic $T(n,\hat{p}_n)$ satisfies* $\mathrm{pr}(T(n,p_0) = T(n,\hat{p}_n)) \rightarrow 1$. ■

As a consequence, the asymptotic properties of $T(n,p_0)$ and $T(n,\hat{p}_n)$ are identical and hence asymptotic inference is unaffected if p_0 is estimated consistently (see Section 4, Remark (iii), however, for a critical discussion of this apparently ideal situation). Although this simple fact has been noted in the literature (Hannan and Quinn [14], p. 191), it seems to have sometimes gone unnoticed as other authors have provided proofs for special instances of this fact (Ensor and Newton [10], Th. 2.1).

Consistency of a number of model selection procedures has been established in different environments in recent years: information criteria like BIC have been shown to provide consistent procedures in the context of stationary autoregressive models (Hannan and Quinn [14], Quinn [27]), stationary autoregressive moving average models (Hannan [12,13]), nonstationary autoregressive models (Paulsen [21], Tsay [35], Paulsen and Tjøstheim [22], Pötscher [25]), and stochastic linear regression models (Geweke and Meese [11], An and Gu [6], Pötscher [25]). Model selection test procedures have also been shown to result in consistent model selection procedures if the significance levels approach zero at an appropriate rate (Pötscher [23], Bauer, Pötscher and Hackl [8], Hosoya [16]), see, for example, Section 3.4, Remark (v).

As a next step we consider the case where a subset $\mathcal{P}_0$ of $\mathcal{P}$ is of special interest for which $\hat{p}_n$ satisfies

$$\operatorname{pr}(\hat{p}_n \in \mathcal{P}_0) \to 1 \quad \text{as} \quad n \to \infty. \tag{1}$$

Typically, $\mathcal{P}_0$ will be the set of all true models, and condition (1) requires that the model selection procedure selects only true, albeit possibly overparametrized, models asymptotically. For example, in the context of estimation of autoregressive models the set $\mathcal{P}_0$ could be chosen as $\{p \in \mathcal{P} : p_0 \leq p\}$, that is, $\mathcal{P}_0$ corresponds to all models in $\mathcal{P}$ which are true but possibly overparametrized, where $\mathcal{P} = \{p \in \mathbb{N}_0 : 0 \leq p \leq P\}$ with $P \geq p_0$ a given upper bound. In this example condition (1) is known to be satisfied for AIC (and a fortiori for consistent procedures like BIC); AIC is furthermore known to overestimate the order p_0 of the autoregressive model asymptotically with positive probability (if $P > p_0$). With $\mathcal{P}_0$ the set of true models, condition (1) is also typically satisfied for model selection procedures based on sequences of tests (Pötscher [23], Bauer, Pötscher and Hackl [8]). In such situations the following "preservation of consistency" result is obviously true.

LEMMA 2. *Let $\mathcal{P}_0$ be finite and assume $T(n,p) \to T$ in probability as $n \to \infty$ for all $p \in \mathcal{P}_0$. If $\hat{p}_n$ and $\mathcal{P}_0$ satisfy* (1), *then $T(n,\hat{p}_n) \to T$ in probability as $n \to \infty$.* (*The range space T, say, of $T(n,p)$ is here assumed to be a metrizable space equipped with its Borel σ-field.*) ■

The significance of Lemma 2 is as follows: If $\mathcal{P}_0$ contains true models only, as is the case in the above example concerning autoregressive models, then many statistics as, for example, parameter estimators based on models $M(p)$, $p \in \mathcal{P}_0$, will converge to the same limit. Lemma 2 then tells us that this limit is also the limit for $T(n,\hat{p}_n)$, provided condition (1) holds. In particular, if $T(n,p)$ represent parameter estimators based on model $M(p)$ which are consistent for one and the same parameter, then consistency carries over to $T(n,\hat{p}_n)$. For example, parameter estimators based on an autoregressive model whose order is estimated by AIC are consistent (of course this is also true as a consequence of Lemma 1 if a consistent model selection procedure like BIC is used). For a further illustration of the implications of Lemma 2 see [26].

3. ASYMPTOTIC DISTRIBUTION

3.1. The Estimation Framework and a Useful Lemma

Lemma 1 shows that the asymptotic distribution of parameter estimators is unaffected by model selection if the model selection procedure is consistent. However, a number of widely used model selection procedures, like procedures based on AIC or C_p, are—although the probability of choosing an incorrect model is asymptotically zero—inconsistent in the sense that the probability of choosing an overparametrized model is asymptotically positive; see Remark (iii) in Section 4 for a discussion of the relative merits of consistent and inconsistent model selection procedures. Hence, it is of interest to ask what the effects of model selection on the asymptotic distribution are in this case. Of course, Lemma 2 does not provide any answer to this question. In this section we study these effects when a particular model selection procedure, which is not consistent and which is described in the next subsection, is used. The estimation framework considered is quite general and can be described as follows: for a sample of size n of the data-generating process we are given an objective function $L_n(\theta)$ from which the estimator is obtained as a minimizer; for example, L_n may be the negative of a log-likelihood function or a sum of squared residuals. The parameter $\theta \in \Theta \subseteq \mathbb{R}^K$ is partitioned as $\theta = (\theta_0', \ldots, \theta_P')'$, where $P \geq 1$ and θ_i is $k_i \times 1$, with $k_i \geq 1$ and $K = k_0 + \cdots + k_P$. The models $M(p)$, $0 \leq p \leq P$, now correspond to the sets $\{\theta \in \Theta : \theta_{p+1} = 0, \ldots, \theta_P = 0\}$. That is, the models $M(p)$ are nested and are obtained from an "overall" model $M(P)$ by imposing the restrictions $\theta_{p+1} = 0, \ldots, \theta_P = 0$. The parameter θ_0 contains parameters common to all models and may contain nuisance parameters. We assume that each $M(p)$ is nonvoid, that is, $(\theta_0', 0, \ldots, 0)' \in \Theta$ holds for some θ_0. We assume further that there is a $\theta^0 \in \Theta$ which is the "true" parameter in a sense made precise below in Assumption A. Let p_0 denote the smallest p, $0 \leq p \leq P$, such that $\theta^0 \in M(p)$. Denote by $\hat{\theta}(p)$ an estimator which minimizes L_n subject to the constraints $\theta_{p+1} = 0, \ldots, \theta_P = 0$ defining model $M(p)$; in particular $\hat{\theta}_i(p) = 0$ for $i > p$. We shall write $\hat{\tau}(p) = (\hat{\theta}_0(p)', \ldots, \hat{\theta}_p(p)')'$, that is, $\hat{\tau}(p)$ consists of the first $K(p) = k_0 + \cdots + k_p$ components of $\hat{\theta}(p)$, and we shall define $\tau^0(p)$ analogously from θ^0. Let $L_{n,p}$ denote the vector of first partial derivatives of L_n w.r.t. $(\theta_0', \ldots, \theta_p')'$ and $L_{n,pp}$ the matrix of corresponding second partial derivatives. We assume $L_n(\theta)$ to be a measurable function on the underlying probability space for each $\theta \in \Theta$. Assumption A below expresses regularity conditions for L_n, with θ^0 playing the role of the "true" parameter. Of course, alternative sets of regularity conditions could be used.

Assumption A. (i) Θ is open in $\mathbb{R}^K$ and L_n has continuous second partial derivatives w.r.t. θ. (ii) $\hat{\theta}(p) \to \theta^0$ in probability as $n \to \infty$, for any $p_0 \leq p \leq P$. (iii) $A(\theta) = \lim_{n\to\infty} n^{-1} E L_{n,PP}(\theta)$ exists and is continuous on Θ. (iv) $A(\theta^0)$ is nonsingular and $n^{-1/2} L_{n,P}(\theta^0)$ converges in distribution to $N(0, A(\theta^0))$. (v) $n^{-1} L_{n,PP}(\theta)$ converges in probability to $A(\theta)$ uniformly in a neighborhood of θ^0, i.e. $\mathrm{pr}(\sup_{\theta \in U(\theta^0)} |n^{-1} L_{n,PP}(\theta) - A(\theta)| > \epsilon) \to 0$ as $n \to \infty$ for some neighborhood $U(\theta^0)$ and all $\epsilon > 0$, where $|\cdot|$ denotes some matrix norm.

The subsequent results also hold with obvious, mainly notational, changes if $k_0 = 0$, that is, if the "nuisance" parameter θ_0 is absent; or if the covariance matrix of the asymptotic distribution of the score $n^{-1/2}L_{n,P}(\theta^0)$ takes the more general form $\nu(\theta^0)A(\theta^0)$ where $\nu(\theta^0)$ is a positive scalar. The latter generalization is of importance if we wish to apply the results, for example, to M-estimators in regression models.

In the following A^0 denotes $A(\theta^0)$ and A_p^0 denotes the leading principal submatrix of A^0 of dimension $K(p)$. The following lemma is the key to the asymptotic distribution of the parameter estimators under model selection given in Theorem 1 below. The lemma may also be of independent interest as it shows that the restricted estimator is asymptotically independent of the unrestricted estimator of those components which are subject to the zero constraints.

LEMMA 3. *Let Assumption A hold. The asymptotic distribution of* $n^{1/2}((\hat{\theta}(p_0) - \theta^0)', \dots, (\hat{\theta}(P) - \theta^0)')'$ *is normal with mean zero and covariance matrix* D^0 *given by* (A5) *in the appendix. Furthermore, for any* p *with* $p_0 \leq p \leq P$ *the set of random variables* $\{\hat{\theta}(r): p_0 \leq r \leq p\}$ *is asymptotically independent of* $\{\hat{\theta}_j(r): p+1 \leq r \leq P, p+1 \leq j\}$. ■

Note that D^0 is singular except in case $p_0 = P$. Of course, Lemma 3 can be appropriately generalized to restrictions more general than zero restrictions, but this is of no concern here.

3.2. The Model Selection Procedure

The model selection procedure considered in the following has been discussed in Anderson [7] in the context of choosing the degree of orthogonal polynomials in polynomial regression. A certain optimality property of the procedure in this and in more general contexts was also established in [7]. For a discussion of the relation of this procedure to information criteria like AIC see Section 3.4, Remark (vi).

At the first stage the procedure consists of testing $M(P-1)$ against $M(P)$, that is, of testing the null hypothesis $\theta_P = 0$ against the alternative $\theta_P \neq 0$. If this test rejects, one puts $\hat{p}_n = P$, otherwise one proceeds with the second stage and tests $M(P-2)$ against $M(P-1)$, that is, one puts $\theta_P = 0$ and tests $\theta_{P-1} = 0$ against $\theta_{P-1} \neq 0$. If the second test now rejects, one puts $\hat{p}_n = P - 1$, otherwise one puts also $\theta_{P-1} = 0$ and proceeds similarly as above. This defines the model selector $\hat{p}_n$ as the largest p, $1 \leq p \leq P$, for which the test of $M(p-1)$ against $M(p)$ rejects, and as 0 if all tests accept. For ease of exposition, we assume in the following that $k_i = 1$ for $i \geq 1$. The test used at each stage consists of rejecting the null hypothesis $M(p-1)$ in favor of $M(p)$, $1 \leq p \leq P$, whenever the test statistic $\hat{t}_p = (\hat{a}_p^0/n)^{-1/2}|\hat{\tau}_p(p)|$ is larger than a critical value c_p. Here $\hat{a}_p^0$ is an estimator of the last diagonal element a_p^0 of $(A_p^0)^{-1}$, that is, of the asymptotic variance of $\hat{\tau}_p(p)$ if $p \geq p_0$. We make the following standard assumption.

Assumption B. (i) $\hat{a}_p^0$ is consistent whenever $p > p_0$. (ii) $\hat{t}_p \to \infty$ in probability as $n \to \infty$, for $1 \le p \le p_0$.

Note that B(ii) is satisfied if $\hat{a}_p^0$ is bounded in probability and $\hat{\tau}_p(p)$ converges in probability to some nonzero value (more generally if $\hat{\tau}_p(p)$ is bounded away from zero in probability) for $1 \le p \le p_0$.

Clearly, under Assumptions A and B, the statistic $\hat{t}_p$ is asymptotically distributed as the absolute value of a standard normal random variable if $p > p_0$, and $\hat{t}_p$ and $\hat{t}_r$ are asymptotically independent in view of Lemma 3 if $p_0 < p < r$.[2] For $1 \le p \le P$ let $0 < \alpha_p < 1$ denote the (nominal) significance level for the test of $M(p-1)$ against $M(p)$; that is, c_p is chosen as the $1 - \alpha_p/2$ quantile of the standard normal distribution. Let $\tilde{\tau}$ denote the parameter estimator which results when the procedure described above is used to select the appropriate model. More formally, $\tilde{\tau} = \hat{\tau}(p)$ if $\hat{p}_n = p$. The corresponding estimator $\tilde{\theta}$ is obtained from $\tilde{\tau}$ by appending the appropriate number of zeroes. Let $\gamma_p = \lim_{n\to\infty} \operatorname{pr}(\hat{p}_n = p)$ for ease of notation.

LEMMA 4. *Under Assumptions A and B, we have* $\gamma_p = 0$ *if* $p < p_0$, $\gamma_{p_0} = (1 - \alpha_{p_0+1}) \cdots (1 - \alpha_P)$, and $\gamma_p = \alpha_p(1 - \alpha_{p+1}) \cdots (1 - \alpha_P)$ *if* $p_0 < p \le P$. ■

With $x \in \mathbb{R}^K$, we use the following notation: $x = (x_0', x_1, \ldots, x_P)'$ where x_0 is $k_0 \times 1$ and x_i, $i > 0$, are scalar. Furthermore $x[p]$ denotes $(x_0', x_1, \ldots, x_p)'$. The cumulative distribution function of a normal random vector with mean zero and covariance matrix Σ is denoted by $\Phi(\cdot;\Sigma)$. We abbreviate $\Phi(\cdot;(A_p^0)^{-1})$ by $\Phi_p(\cdot)$. The corresponding densities are denoted by $\phi(\cdot;\Sigma)$ and $\phi_p(\cdot)$, respectively. The main result now gives the asymptotic distribution of $n^{1/2}(\tilde{\theta} - \theta^0)$.[3]

THEOREM 1. *Let Assumptions A and B hold. For* $x \in \mathbb{R}^K$, *we have*

$$\lim_{n\to\infty} \operatorname{pr}(n^{1/2}(\tilde{\theta} - \theta^0) \le x) = \sum_{p=p_0}^{P} \gamma_p F_p(x)$$

where F_p *is given as follows*: $F_{p_0}(x) = \Phi_{p_0}(x[p_0])$ *if* $x_i \ge 0$ *for* $i > p_0$, *and* $F_{p_0}(x) = 0$ *otherwise. For* $p_0 < p \le P$, *we have*

$$F_p(x) = \alpha_p^{-1}\{\Phi_p(x[p-1], \min(x_p, -(a_p^0)^{1/2}c_p)) + \max\{0, \Phi_p(x[p]) - \Phi_p(x[p-1], (a_p^0)^{1/2}c_p)\}\}$$

if $x_i \ge 0$ *for* $i > p$, *and* $F_p(x) = 0$ *otherwise.*[4] *Moreover,* $F_p(x)$ *is the limit of the conditional distribution* $\operatorname{pr}(n^{1/2}(\tilde{\theta} - \theta^0) \le x \mid \hat{p}_n = p)$ *for* $p \ge p_0$. ■

Note that the convergence of the distribution function as well as of the conditional distribution functions in Theorem 1 is for all $x \in \mathbb{R}^K$ and not only for continuity points of the limiting distribution. The limiting conditional distribution F_p is clearly concentrated on the subspace defined by $x_i = 0$ for all $i > p$, and hence is not absolutely continuous on $\mathbb{R}^K$, due to the fact that some of the components of the estimator are restricted to zero on the conditioning event. However, if we view F_p as a distribution on this subspace, that is, if we disregard the components restricted to zero, then it has a density $f_p(x[p])$. (In other words, f_p is the density of the limiting distribution of $\tilde{\tau}$ conditional on $\{\hat{p}_n = p\}$.) From Theorem 1 we get the following simple expressions for these densities: $f_{p_0}(x[p_0]) = \phi_{p_0}(x[p_0])$ and, for $p_0 < p \leq P$,

$$f_p(x[p]) = \begin{cases} \alpha_p^{-1}\phi_p(x[p]) & \text{for } |x_p| > (a_p^0)^{1/2}c_p \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

We next discuss shortly some of the implications of Theorem 1 for econometric and statistical methodology. (A more detailed discussion of these implications and of aspects of the practical implementation will be given in a subsequent paper.) Theorem 1 puts us in the position to draw valid inferences albeit using the same data set for model selection and estimation. Suppose, for example, we want to construct a confidence region for θ^0 conditional on the event that model $M(p)$ has been chosen by the model selection procedure, that is, conditional on the event $\{\hat{p}_n = p\}$. Assuming for the moment that the covariance matrix A^0 is known, we can then calculate the confidence region from the formulae for F_p or f_p. However, since these formulae are different depending on whether $(\hat{p}_n =)p = p_0$ or $(\hat{p}_n =)p > p_0$ holds and since p_0 is unknown, we have to calculate two confidence regions, one using the formula on the presumption $p > p_0$ and one on the presumption $p = p_0$. Note that the latter confidence region is, since $f_{p_0} = \phi_{p_0}$ holds, precisely the "classical" confidence region one would use if the model $M(p)$ had been chosen a priori without reference to the data set. Hence, we would report two confidence regions, one of the two being the valid confidence region (but we would not know which one is which). Of course, another way of providing information on the variability of parameter estimators is to report the (asymptotic conditional) covariance matrices, which can be calculated analytically from the formulae for F_p or f_p. Again two such covariance matrices would have to be reported corresponding to the cases $(\hat{p}_n =)p = p_0$ and $(\hat{p}_n =)p > p_0$. Also upper and lower bounds for these matrices, similar to the ones given in the corollary below, can be derived and can be used in assessing variability of the estimators. A further important implication, which can be gleaned from Theorem 1, concerns the asymptotic conditional distribution of $\tilde{\theta}$ given that the model selection procedure has selected the model $M(p_0)$, that is, the smallest true model in the model class considered. Theorem 1 shows that this distribution is identical to the classical asymptotic distribution of $\hat{\theta}(p_0)$, that is, of the parameter estimator based on the a priori chosen model $M(p_0)$. (Note, however, that the probability of the event $\{\hat{p}_n = p_0\}$ may be substantially less than 1, and that we do not of course know whether this event does or does not hold in a particular application.)

Since F_p and f_p depend on A^0, which is usually unknown, we have to use $\hat{F}_p$ and $\hat{f}_p$ in practice where $\hat{F}_p$ and $\hat{f}_p$ are obtained from F_p and f_p by replacing A_p^0 and a_p^0 with consistent estimators in the respective formulae. The resulting distribution function $\hat{F}_p$ is a valid approximation to $\text{pr}(n^{1/2}(\tilde{\theta} - \theta^0) \leq x \mid \hat{p}_n = p)$ since it is not difficult to establish that $\sup_x |\hat{F}_p(x) - F_p(x)| \to 0$ as the sample size increases. Hence, $\hat{F}_p$ approximates the conditional distribution $\text{pr}(n^{1/2}(\tilde{\theta} - \theta^0) \leq x \mid \hat{p}_n = p)$ in exactly the same fashion as F_p does. The same is of course also true for the unconditional distribution. Note that in the important special case where $A^0 = \sigma^{-2}(\theta^0)Q$, Q independent of the parameter, standardizing the estimator by $\sigma(\theta^0)$ gives the asymptotic conditional distribution in a form independent of unknown parameters. This shows that estimating the unknown quantities in the asymptotic distributions as suggested above has—at least in this special case—no more adverse an effect on the quality of approximation than in a classical asymptotic situation.

For the purpose of illustration of the effect of model selection on the asymptotic distribution of the parameter estimators, we next give as a corollary the asymptotic distribution of $\tilde{\delta} = (\tilde{\theta}_0', \ldots, \tilde{\theta}_{p_0})'$, that is, of that subvector of the estimator $\tilde{\theta}$ corresponding to the parameters in the minimal true model, in a form more suitable for comparison with the asymptotic distribution of the corresponding estimator based on model $M(p)$, p fixed. (Of course, the limiting conditional distribution of other subvectors of $\tilde{\theta}$ can also be expressed in a similar form.) Furthermore, the asymptotic covariance matrix of $\tilde{\delta}$ is shown to be bounded by the asymptotic covariance matrices $B_{p_0}^0$ and B_P^0 of the corresponding estimators based on $M(p_0)$ and $M(P)$, respectively. We first introduce some notation: for $p \geq p_0$, let B_p^0 denote the leading principal $K(p_0) \times K(p_0)$ submatrix of $(A_p^0)^{-1}$. For $p > p_0$, define V_p^0 and W_p^0 via

$$A_p^0 = \begin{bmatrix} A_{p_0}^0 & W_p^{0\prime} \\ W_p^0 & V_p^0 \end{bmatrix},$$

put $\mu_p = (0,\ldots,0,1)(V_p^0)^{-1}W_p^0$, and $\sigma_p^2 = (0,\ldots,0,1)(V_p^0)^{-1}(0,\ldots,0,1)'$. Let Φ^* denote the standard normal distribution function.

COROLLARY. *Let Assumptions A and B hold. For $p \geq p_0$, the limit of the conditional distribution* $\text{pr}(n^{1/2}(\tilde{\delta} - \delta^0) \leq z \mid \hat{p}_n = p)$ *has a density g_p given by*

$$g_p(z) = \kappa_p(z,\alpha_p)\phi(z;B_p^0)$$

with

$$\kappa_p(z,\alpha_p) = \alpha_p^{-1}\{2 - \Phi^*(\sigma_p^{-1}[(a_p^0)^{1/2}c_p - \mu_p z]) - \Phi^*(\sigma_p^{-1}[(a_p^0)^{1/2}c_p + \mu_p z])\}$$

if $p > p_0$, and $\kappa_{p_0}(z,\alpha_{p_0}) = 1$. For $p > p_0$, the covariance matrix of g_p, M_p^0 say, satisfies $B_p^0 \leq M_p^0 \leq \alpha_p^{-1}B_p^0$. Furthermore, the asymptotic distribution of $n^{1/2}(\tilde{\delta} - \delta^0)$ has a density g given by $g(z) = \sum_{p=p_0}^{P} \gamma_p g_p(z)$. The (uncon-

ditional) asymptotic covariance matrix, M^0 say, of $n^{1/2}(\tilde{\delta} - \delta^0)$ satisfies $B^0_{p_0} \leq M^0 \leq B^0_P$.[5] ■

The corollary shows that the density g_p of the limiting conditional distribution differs from the density $\phi(z;B^0_p)$ of the distribution of the corresponding estimator based on model $M(p)$, p fixed, only by the scalar factor $\kappa_p(z,\alpha_p)$. As is easily seen, $\kappa_p(z,\alpha_p) \leq \alpha_p^{-1}$ holds for $p > p_0$; furthermore, κ_p and hence g_p are symmetric about zero. Of course, $B^0_{p_0} = (A^0_{p_0})^{-1}$ implies $g_{p_0} = \phi_{p_0}$ and hence $M^0_{p_0} = B^0_{p_0}$ (in accordance with the discussion after Theorem 1).

3.3. Numerical Examples

Consider a distributed lag model of the form

$$y_t = \sum_{i=0}^{3} \theta_i x_{t-i} + \epsilon_t$$

where we assume that (x_t) follows either a stationary AR(1) or MA(1) process and is independent of (ϵ_t), which for simplicity is assumed to be i.i.d. with zero mean and finite variance (as the OLS estimator of the variance of ϵ_t is asymptotically independent of the OLS estimator of the parameter vector θ we may and do assume that the variance is known to be unity without affecting the results). In terms of Section 3.1 we have $P = 3$, $k_0 = 1$ and $-L_n$ is the Gaussian (quasi) log-likelihood. We further assume that the true model does not contain lags, that is, $\theta^0 = (\theta^0_0,0,0,0)$ and hence $p_0 = 0$. If (x_t) follows an AR(1)-model, that is, $x_t = \rho x_{t-1} + u_t$, $|\rho| < 1$, $Eu_t^2 = 1$, (u_t) i.i.d., then the matrix $A(\theta^0)$ is the 4×4 Toeplitz matrix with elements $(1 - \rho^2)^{-1}\rho^{|i-j|}$. If (x_t) follows an MA(1)-model, that is, $x_t = u_t + \rho u_{t-1}$, $|\rho| < 1$, $Eu_t^2 = 1$, (u_t) i.i.d., then the matrix $A(\theta^0)$ is the 4×4 Toeplitz matrix with elements $1 + \rho^2$ on the main diagonal, ρ on the adjacent diagonals, and zeroes elsewhere. Of course, the same matrices $A(\theta^0)$ may also arise from another, even nonlinear, model, and hence the numerical results presented below also hold for such models. The distributed lag model is only chosen as a vehicle to generate reasonable matrices $A(\theta^0)$.

To illustrate the effects of model selection on the asymptotic distribution we compare the actual limiting (conditional) density $g_p(z)$ for the estimator $\tilde{\theta}_0$ with the density $\phi(z;B^0_p)$ of the asymptotic distribution of $\hat{\theta}_0(p)$, that is, with the density if p is held fixed (that is, with the nominal density one would use if one ignores randomness of $\hat{p}_n$). Both densities, $g_p(z)$ and $\phi(z;B^0_p)$ were calculated for the AR(1) as well as for the MA(1) case for $\rho = 0.3$, 0.5, 0.8, 0.9, $\alpha_p = 0.01$, 0.05, 0.1 and $p = 1,2,3$. (As the results are invariant under $\rho \to -\rho$, no results for negative ρ are reported below.) The differences between $g_p(z)$ and $\phi(z;B^0_p)$ are quite substantial. Graphs of both densities for selected parameter combinations are shown in Figures 1–6. Since $g_p(z)$

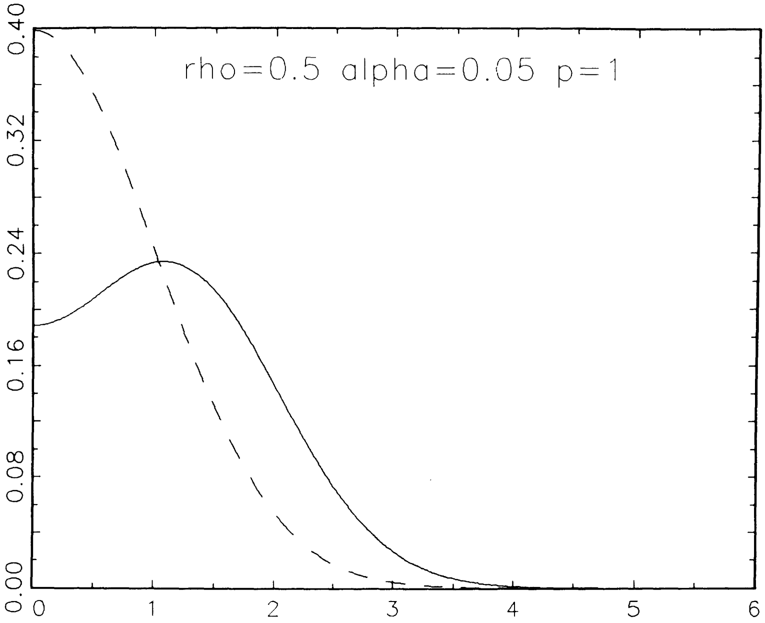

Figure 1. g_p solid line, $\phi(\,.\,;B_p^0)$ broken line. AR(1)-regressors.

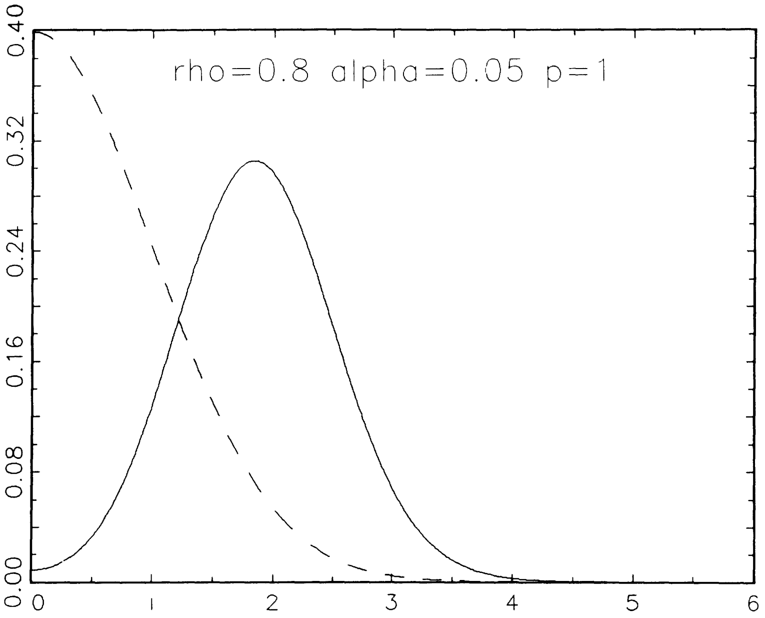

Figure 2. g_p solid line, $\phi(\,.\,;B_p^0)$ broken line. AR(1)-regressors.

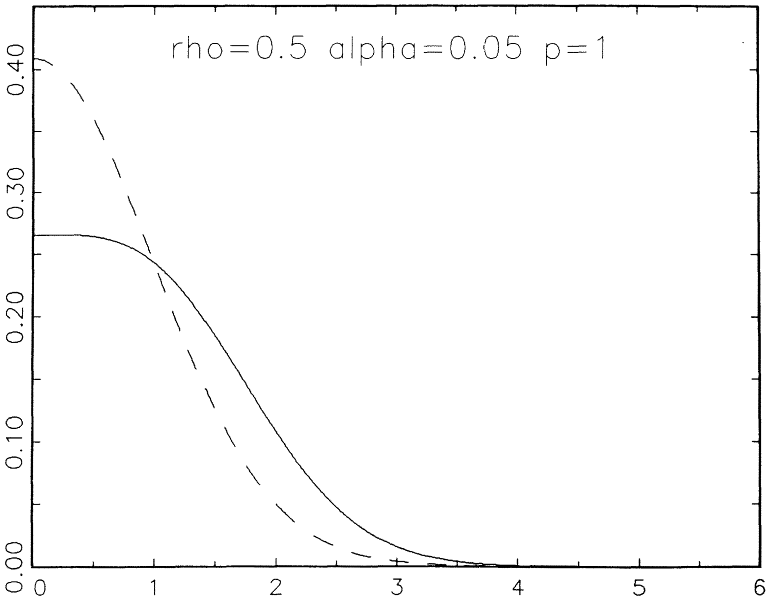

FIGURE 3. g_p solid line, $\phi(\,.\,;B_p^0)$ broken line. MA(1)-regressors.

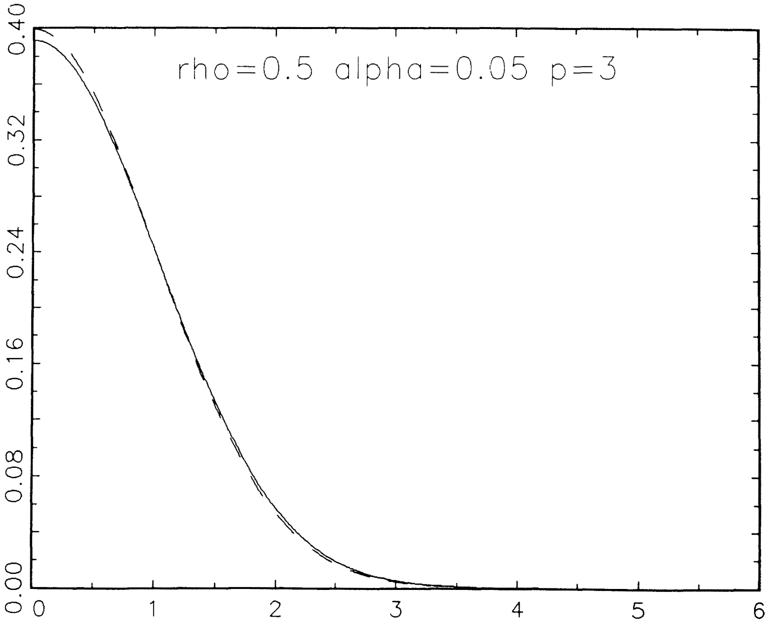

FIGURE 4. g_p solid line, $\phi(\,.\,;B_p^0)$ broken line. MA(1)-regressors.

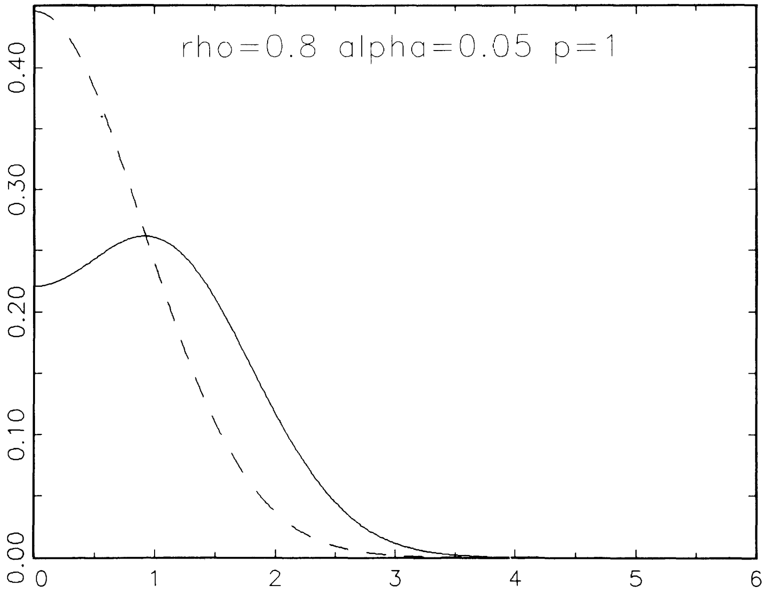

FIGURE 5. g_p solid line, $\phi(\,.\,;B_p^0)$ broken line. MA(1)-regressors.

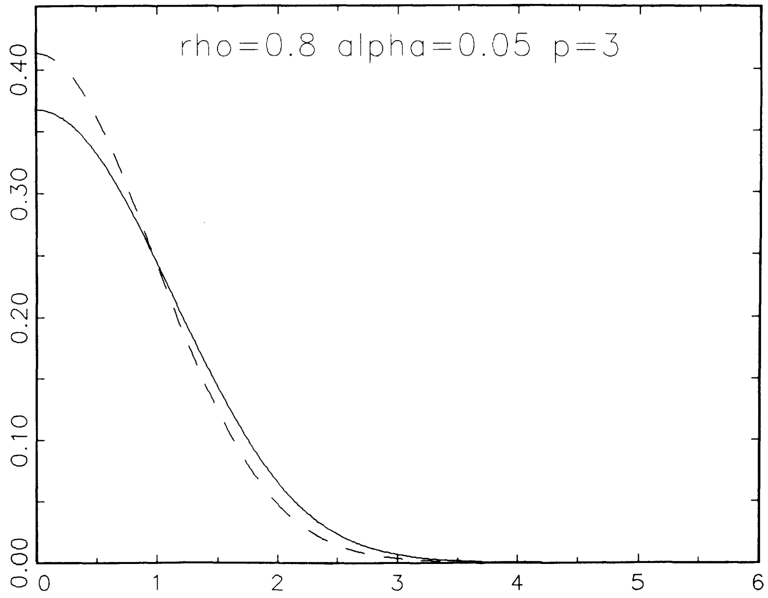

FIGURE 6. g_p solid line, $\phi(\,.\,;B_p^0)$ broken line. MA(1)-regressors.

and $\phi(z;B_p^0)$ are both symmetric about zero, the values of both densities are plotted only for nonnegative values of z. As is to be expected, g_p is more spread out compared with $\phi(\cdot;B_p^0)$. In fact, g_p becomes bimodal in some cases. In general, the difference between the two densities becomes larger as $|\rho|$ increases. The difference is also more pronounced the smaller α_p is (note, however, that smaller values of α_p in the range considered imply smaller probabilities of the event $\{\hat{p}_n = p\}$, $p > p_0 = 0$). Furthermore, due to the structure of the covariance matrix of an AR(1) (MA(1)) process, the entries of $A(\theta^0)$ decrease as one moves away from the main diagonal, hence the difference between $g_p(z)$ and $\phi(z;B_p^0)$ decreases with increasing p; this feature is, however, due to the special structure of the chosen matrix $A(\theta^0)$. From theoretical considerations it is furthermore clear that g_p and $\phi(z;B_p^0)$ must coincide in the AR-case if $p > 1$ and hence no results are presented for this case. (Note that in the AR(1) case $A(\theta^0)^{-1}$ is a band matrix with bandwidth 3, hence the tests leading to $\hat{p}_n = p$, $p > 1$, are asymptotically independent of the estimator for θ_0^0. As a program check these cases were also calculated and the expected result was obtained.) Because of space limitations we do not give the graphs for $p = 2$ in the MA-case; the corresponding densities show a behavior intermediate to the cases $p = 1$ and $p = 3$.

To illustrate further the impact of model selection on inference, we compare the correct confidence interval based on g_p with the nominal confidence interval based on $\phi(z;B_p^0)$. Since both densities are symmetric about zero, the right-hand endpoint of the $1 - \beta$ confidence interval (shifted by the true parameter value θ_0^0 and scaled by $n^{1/2}$) is given by the $1 - \beta/2$ quantile of the corresponding density. These quantiles, denoted by $c(g)$ and $c(\phi)$, respectively, are reported in Tables 1–4 for $\beta = 0.5, 0.2, 0.1$, and for selected parameter combinations. It transpires that the nominal confidence intervals based on classical asymptotic theory can be quite misleading, as the true confidence intervals may be larger up to a factor of 3.4. For example, if the regressors follow a MA(1) process with $\rho = 0.8$ and $\alpha_p = 0.1$, we see from Table 2 that the correct 90% confidence interval given $\hat{p}_n = 1$ (2,3) is 31% (15%,8%) longer compared with the nominal 90% confidence interval obtained from "standard" asymptotics ignoring randomness of $\hat{p}_n$. Note that the (limiting) probability of the event $\{\hat{p}_n > p_0\}$ is as large as 0.27 in this case.

3.4. Remarks

i. In some cases, like maximum likelihood estimation of the standard normal linear regression model, the finite sample distribution and the asymptotic distribution coincide, and then Lemma 3 holds in finite samples.

ii. A heuristic derivation of Lemma 3 in the context of maximum likelihood estimation is as follows: for simplicity assume $P = 1$, $k_0 = k_1 = 1$

TABLE 1. Right-hand endpoints $c(g)$ and $c(\phi)$ of $1 - \beta$ confidence intervals for θ_0^0 based on g_p and $\phi(\,.\,;B_p^0)$, respectively (MA(1) regressors)

$\rho = 0.5$				
α_p	p	β	$c(g)$	$c(g)/c(\phi)$
0.10	1	0.5	0.876	1.331
		0.2	1.581	1.264
		0.1	1.972	1.229
	2	0.5	0.714	1.065
		0.2	1.353	1.062
		0.1	1.733	1.060
	3	0.5	0.684	1.015
		0.2	1.299	1.015
		0.1	1.667	1.015
0.05	1	0.5	0.956	1.452
		0.2	1.680	1.343
		0.1	2.074	1.292
	2	0.5	0.729	1.088
		0.2	1.380	1.083
		0.1	1.766	1.080
	3	0.5	0.687	1.021
		0.2	1.306	1.020
		0.1	1.676	1.020
0.01	1	0.5	1.142	1.734
		0.2	1.889	1.510
		0.1	2.285	1.424
	2	0.5	0.768	1.145
		0.2	1.445	1.134
		0.1	1.842	1.126
	3	0.5	0.696	1.034
		0.2	1.322	1.033
		0.1	1.696	1.033

and $p_0 = 0$. Then $\hat{\theta}_1(1) \to 0$. Now for any $\lambda \in \mathbb{R}$, the estimator $\tilde{\vartheta}(\lambda) = \hat{\theta}_0(0) + \lambda\hat{\theta}_1(1)$ is consistent for $\theta_0(0)$. If $\hat{\theta}_0(0)$ and $\hat{\theta}_1(1)$ would have nonzero asymptotic correlation, then for some value of λ, the estimator $\tilde{\vartheta}(\lambda)$ would have a smaller asymptotic variance than $\hat{\theta}_0(0)$, thus contradicting asymptotic efficiency of the restricted maximum likelihood estimator.

iii. Note that p_0 is defined relative to θ^0, and that uniqueness of θ^0 follows from Assumption A(ii) given a particular choice for the estimator sequence $\hat{\theta}(p)$. Hence, if L_n has multiple minimizers, θ^0 would in

Table 2. Right-hand endpoints $c(g)$ and $c(\phi)$ of $1-\beta$ confidence intervals for θ_0^0 based on g_p and $\phi(\,.\,;B_p^0)$, respectively (MA(1) regressors)

$\rho = 0.8$

α_p	p	β	$c(g)$	$c(g)/c(\phi)$
0.10	1	0.5	0.920	1.526
		0.2	1.573	1.372
		0.1	1.924	1.307
	2	0.5	0.755	1.188
		0.2	1.406	1.166
		0.1	1.782	1.151
	3	0.5	0.709	1.089
		0.2	1.341	1.084
		0.1	1.715	1.080
0.05	1	0.5	1.028	1.705
		0.2	1.689	1.474
		0.1	2.040	1.386
	2	0.5	0.799	1.257
		0.2	1.472	1.219
		0.1	1.853	1.197
	3	0.5	0.730	1.121
		0.2	1.377	1.113
		0.1	1.757	1.106
0.01	1	0.5	1.262	2.092
		0.2	1.927	1.681
		0.1	2.276	1.547
	2	0.5	0.907	1.428
		0.2	1.616	1.339
		0.1	2.006	1.295
	3	0.5	0.782	1.201
		0.2	1.460	1.180
		0.1	1.852	1.166

principle only be identified by the chosen estimator. However, in most cases of interest the limit of the estimator sequence does not depend on the particular choice for $\hat{\theta}(p)$. For example, if L_n is the negative of a log-likelihood, then θ^0 is indeed the true parameter and—under regularity conditions—any sequence of minimizers converges to the same limit θ^0.

iv. We note that Assumption A rules out order estimation in autoregressive moving average models due to identifiability problems (Hannan [12,13], Pötscher [23,24]).

TABLE 3. Right-hand endpoints $c(g)$ and $c(\phi)$ of $1-\beta$ confidence intervals for θ_0^0 based on g_p and $\phi(\,.\,;B_p^0)$, respectively (AR(1) regressors)

$\rho = 0.5$				
α_p	p	β	$c(g)$	$c(g)/c(\phi)$
0.10	1	0.5	1.050	1.556
		0.2	1.778	1.387
		0.1	2.168	1.318
0.05	1	0.5	1.176	1.743
		0.2	1.912	1.492
		0.1	2.301	1.399
0.01	1	0.5	1.446	2.144
		0.2	2.185	1.705
		0.1	2.572	1.564

TABLE 4. Right-hand endpoints $c(g)$ and $c(\phi)$ of $1-\beta$ confidence intervals for θ_0^0 based on g_p and $\phi(\,.\,;B_p^0)$, respectively (AR(1) regressors)

$\rho = 0.8$				
α_p	p	β	$c(g)$	$c(g)/c(\phi)$
0.10	1	0.5	1.637	2.427
		0.2	2.205	1.720
		0.1	2.512	1.527
0.05	1	0.5	1.859	2.756
		0.2	2.418	1.887
		0.1	2.718	1.653
0.01	1	0.5	2.306	3.419
		0.2	2.850	2.224
		0.1	3.141	1.910

v. If the significance levels α_p are allowed to depend on sample size and if they converge to zero at an appropriate rate as sample size increases, then the model selector $\hat{p}_n$ considered in Section 3.2 is consistent for p_0, see Pötscher [23], and Bauer, Pötscher, and Hackl [8].

vi. Under standard regularity conditions, the square of the test statistic $\hat{t}_p$ is asymptotically equivalent to the test statistic $s_p = -2(L_n(\hat{\theta}(p)) - L_n(\hat{\theta}(p-1)))$, that is, to the likelihood ratio test statistic if $-L_n$ is a log-likelihood. In particular, Lemma 3 implies asymptotic independence of the test statistics s_p for $p > p_0$. Further, Theorem 1 also applies to the model selection procedure which uses s_p instead of $\hat{t}_p$. If we additionally choose the significance levels α_p such that the corresponding critical values $\bar{c}_p$, that is, the $1 - \alpha_p$ quantiles of a chi-square variable with one degree of freedom, satisfy $\bar{c}_p = 2$, then this new procedure is identical with the following modification of the minimum AIC method: let $\text{AIC}(p) = 2L_n(\hat{\theta}(p)) + 2(k_0 + p)$ and define $\hat{p}_n$ to be the largest of $P, P - 1, \ldots$ for which AIC has a "local" minimum, that is, for which $\text{AIC}(p) \geq \text{AIC}(\hat{p}_n)$ for $p > \hat{p}_n$, and $\text{AIC}(\hat{p}_n) < \text{AIC}(\hat{p}_n - 1)$. Compare also Remark (ii) in Section 4.

vii. The results of Section 3.2 immediately give the asymptotic distribution of the usual predictors in a linear regression model with nonstochastic regressors, as these predictors are only linear combinations of the parameter estimators.

4. SOME COMMENTS ON MODEL SELECTION

i. As mentioned in Section 2, parameter estimators, derived from autoregressive models whose order is selected by a procedure like AIC which is not consistent but satisfies (1), are consistent as a consequence of Lemma 2. This phenomenon is generally true if all models are contained in an overall model and if the parameters of the smaller models remain identifiable in the overall model. However, if identifiability of these parameters is lost in the overall model, as is, for example, the case with autoregressive moving average models, then the above statement is no longer true. In such "nonregular" cases consistent model selection becomes especially important, if the focus is on estimation of these parameters.

ii. Minimizing a criterion like AIC or BIC amounts to testing any model against all other models by means of likelihood ratio tests (if $-L_n$ is a log-likelihood) and to select that model which is accepted against all other models. The critical values are determined by the penalty terms of the criteria. For further discussion see Akaike [4] and Söderström [34]. See also Remark (vi) in the preceding section.

iii. Suppose that an overall model exists, that the submodels are described by lower-dimensional submanifolds of Θ, and that the parameter vector θ is identified in the overall model. Then consistent model selection procedures followed by (quasi) maximum likelihood estimation of the parameters of the selected model typically lead to superefficient estimators of θ in view of Lemma 1. It should be noted, however, that the risk of superefficient estimators near the true parameter value has unpleasant properties (Lehmann [19], p. 407). Another unpleasant aspect of this superefficiency is the fact that the convergence of the finite sample distribution to the asymptotic distribution is not uniform near the lower-dimensional submanifolds. Of course, this is the price to be paid for superefficiency. For this reason Shibata [33] favors AIC over BIC if not only estimation of the model order but also parameter estimation is of concern (and identifiability is not lost in overparametrized models). See also Shibata [30,32] and Hosoya [15] for a discussion of the risk involved in using model selection procedures like AIC or BIC. (To a lesser extent the same unpleasant effects also arise with AIC or with the procedure of Section 3.) Of course, if estimation of p_0 is the goal rather than estimation of θ, then consistent procedures are preferable. For "nonregular" cases, where identifiability is lost in the overall model, the above discussion does not directly apply. Furthermore, in such a case consistent estimation of p_0 becomes especially important, if one wants to select a model in which all parameters are identified.

NOTES

[1] The closely related literature on pretest estimators concentrates mainly on first and second moments of the finite sample distributions of the estimators, see [17,18]. An alternative approach to the distribution of pretest estimators than the one presented here is via bootstrapping, see [9] and the references therein.

[2] It is interesting to note that the distribution of $\hat{t}_p$ conditional on the event that $M(p)$, $M(p+1), \ldots$ have been accepted at the previous stages is identical to the unconditional distribution in view of the independence property mentioned.

[3] After the original version of the paper was completed, Don Andrews brought the paper by Sen [29] to my attention. In [29] the special case corresponding to $p_0 = 0$, $P = 1$ is treated in an i.i.d. maximum likelihood context.

[4] Note that the form of the limiting distribution F_p depends effectively only on the test of $M(p-1)$ against $M(p)$, in particular F_p is independent of $\alpha_{p+1}, \ldots, \alpha_P$.

[5] If, furthermore, $B^0_{p_0} < B^0_p$ for some p, $p_0 < p \le P$, holds then even $B^0_{p_0} < M^0$ is true, as is easily seen. If $B^0_p < B^0_P$ for some p, $p_0 \le p < P$, holds, then $M^0 < \eta B^0_P$ with $\eta = \gamma_{p_0} + \sum_{p=p_0+1}^{P} (\gamma_p/\alpha_p)$ follows using the upper bounds for M^0_p. (I have not been able to establish this *strict* inequality without the factor η.)

REFERENCES

1. Akaike, H. Fitting autoregressive models for prediction. *Annals of the Institute of Statistical Mathematics* 21 (1969): 243-247.
2. Akaike, H. Statistical predictor identification. *Annals of the Institute of Statistical Mathematics* 22 (1970): 203-217.
3. Akaike, H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 19 (1974): 716-723.
4. Akaike, H. Comments on "On model structure testing in system identification." *International Journal of Control* 27 (1977): 323-324.
5. Amemiya, T. Selection of regressors. *International Economic Review* 21 (1980): 331-354.
6. An, H.Z. & L. Gu. On selection of regression variables. *Acta Mathematicae Applicandae Sinica* 2 (1985): 27-36.
7. Anderson, T.W. The choice of the degree of a polynomial regression as a multiple decision problem. *Annals of Mathematical Statistics* 33 (1962): 255-265.
8. Bauer, P., B.M. Pötscher & P. Hackl. Model selection by multiple test procedures. *Statistics* 19 (1988): 39-44.
9. Brownstone, D. How to "data mine" if you must: bootstrapping Stein-rule model selection

procedures. Technical Report MBS 90-08, Irvine Research Unit in Mathematical Behavioral Sciences, UC Irvine, 1990.

10. Ensor, K.B. & H.J. Newton. The effect of order estimation on estimating the peak frequency of an autoregressive spectral density. *Biometrika* 75 (1988): 587-589.
11. Geweke, J. & R. Meese. Estimating regression models of finite but unknown order. *International Economic Review* 22 (1981): 55-70.
12. Hannan, E.J. The estimation of the order of an ARMA process. *Annals of Statistics* 8 (1980): 1071-1081.
13. Hannan, E.J. Estimating the dimension of a linear system. *Journal of Multivariate Analysis* 11 (1981): 459-473.
14. Hannan, E.J. & B.G. Quinn. The determination of the order of an autoregression. *Journal of the Royal Statistical Society* B 41 (1979): 190-195.
15. Hosoya, Y. Information criteria and tests for time series models. In O.D. Anderson (ed.), *Time Series Analysis: Theory and Practice 5*, pp. 39-52. Amsterdam: North-Holland, 1984.
16. Hosoya, Y. Hierarchical statistical models and a generalized likelihood ratio test. *Journal of the Royal Statistical Society* B 51 (1989): 435-447.
17. Judge, G.G. & M.E. Bock. *The Statistical Implications of Pre-Test and Stein-Rule Estimators in Econometrics.* Amsterdam: North-Holland, 1978.
18. Judge, G.G. & T.A. Yancey. *Improved Methods of Inference in Econometrics.* Amsterdam: North-Holland, 1986.
19. Lehmann, E.L. *Theory of Point Estimation.* New York: Wiley, 1983.
20. Mallows, C.L. Some comments on C_p. *Technometrics* 15 (1973): 661-675.
21. Paulsen, J. Order determination of multivariate autoregressive time series with unit roots. *Journal of Time Series Analysis* 5 (1984): 115-127.
22. Paulsen, J. & D. Tjøstheim. Least squares estimates and order determination procedures for autoregressive processes with a time dependent variance. *Journal of Time Series Analysis* 6 (1985): 117-133.
23. Pötscher, B.M. Order estimation in ARMA-models by Lagrangian multiplier tests. *Annals of Statistics* 11 (1983): 872-885.
24. Pötscher, B.M. The behaviour of the Lagrangian multiplier test in testing the orders of an ARMA-model. *Metrika* 32 (1985): 129-150.
25. Pötscher, B.M. Model selection under nonstationarity: autoregressive models and stochastic linear regression models. *Annals of Statistics* 17 (1989): 1257-1274.
26. Pötscher, B.M. Effects of model selection on inference. Working Paper, Institut für Ökonometrie, Operations Research und Systemtheorie, Technische Universität Wien, 1989.
27. Quinn, B.G. Order determination for multivariate autoregression. *Journal of the Royal Statistical Society* B 42 (1980): 182-185.
28. Schwarz, G. Estimating the dimension of a model. *Annals of Statistics* 6 (1978): 461-464.
29. Sen, P.K. Asymptotic properties of maximum likelihood estimators based on conditional specification. *Annals of Statistics* 7 (1979): 1019-1033.
30. Shibata, R. Selection of the order of an autoregressive model by Akaike's information criterion. *Biometrika* 63 (1976): 117-126.
31. Shibata, R. Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. *Annals of Statistics* 8 (1980): 147-164.
32. Shibata, R. A theoretical view of the use of AIC. In O.D. Anderson (ed.), *Time Series Analysis: Theory and Practice 4*, pp. 237-244. Amsterdam: North-Holland, 1983.
33. Shibata, R. Consistency of model selection and parameter estimation. In J. Gani and M.B. Priestley (eds.), *Essays in Time Series and Allied Processes*, pp. 127-141. Sheffield: Applied Probability Trust, 1986.
34. Söderström, T. On model structure testing in system identification. *International Journal of Control* 26 (1977): 1-18.
35. Tsay, R.S. Order selection in nonstationary autoregressive models. *Annals of Statistics* 12 (1984): 1425-1433.

APPENDIX

Proof of Lemma 3. Because of the definition of p_0 and of $\hat{\theta}(r)$ for $p_0 \leq r \leq P$ we have in view of Assumption A and the mean value theorem

$$O = L_{n,r}(\hat{\theta}(r)) = L^0_{n,r} + L^*_{n,rr}(\hat{\tau}(r) - \tau^0(r)) \tag{A1}$$

where $L^0_{n,r}$ is $L_{n,r}(\theta^0)$ and $L^*_{n,rr}$ is $L_{n,rr}$ evaluated row-wise at mean values lying on the line segment joining $\hat{\theta}(r)$ and θ^0. Observe that

$$(L^{0\prime}_{n,p_0},\ldots,L^{0\prime}_{n,P})' = J(p_0,P)L^0_{n,P} \tag{A2}$$

where

$$J(p_0,P) = \begin{bmatrix} I(K(p_0)) & O(K(p_0),K(P) - K(p_0)) \\ I(K(p_0+1)) & O(K(p_0+1),K(P) - K(p_0+1)) \\ \vdots & \vdots \\ I(K(P)) & \end{bmatrix}$$

is $(K(p_0)+\cdots+K(P)) \times K(P)$, $I(m)$ is the $m \times m$ identity matrix, and $O(m,M-m)$ is a zero matrix of dimension $m \times (M-m)$. Now multiplying (A1) by $n^{-1/2}$ and stacking (A1) for $p_0 \leq r \leq P$, we obtain from (A2)

$$n^{-1/2}J(p_0,P)L^0_{n,P}$$
$$= -n^{-1}\operatorname{diag}(L^*_{n,p_0p_0},\ldots,L^*_{n,PP})n^{1/2}((\hat{\tau}(p_0) - \tau^0(p_0))',\ldots,(\hat{\tau}(P) - \tau^0(P))')'. \tag{A3}$$

As a consequence of Assumption A, $n^{-1/2}J(p_0,P)L^0_{n,P}$ is asymptotically normal with mean zero and covariance matrix $J(p_0,P)A^0J(p_0,P)'$ and $n^{-1}\operatorname{diag}(L^*_{n,p_0p_0},\ldots,L^*_{n,PP})$ converges in probability to $E = \operatorname{diag}(A^0_{p_0},\ldots,A^0_P)$ which is nonsingular as a consequence of Assumption A. (If $L^*_{n,rr}$ is not measurable, then convergence in probability has to be formulated in terms of outer probability.) Hence, $T = n^{1/2}((\hat{\tau}(p_0) - \tau^0(p_0))',\ldots,(\hat{\tau}(P) - \tau^0(P))')'$ is asymptotically normal with mean zero and covariance matrix $C^0 = E^{-1}J(p_0,P)A^0J(p_0,P)'E^{-1}$ because of (A3). Now partitioning C^0 conformably with T, simple calculation shows that for $u \leq v$, the (u,v)th block C^0_{uv} of C^0 is given by

$$C^0_{uv} = [(A^0_{p_0+u-1})^{-1} : O(K(p_0+u-1),K(p_0+v-1) - K(p_0+u-1))]. \tag{A4}$$

To prove the second half of the lemma, observe that it is clearly true if $p = P$, since then the second set in the formulation of the lemma is void. Hence, assume $p_0 \leq p < P$. Consider an element $\hat{\theta}(r)$ of the first set, that is, $p_0 \leq r \leq p$, and an element $\hat{\theta}_j(s)$ of the second set, that is, $p+1 \leq s \leq P$ and $p+1 \leq j$ hold. Then $\hat{\theta}_i(r) = \hat{\tau}_i(r)$ for $0 \leq i \leq r$ and $\hat{\theta}_i(r) = 0$ for $i > r$. Similarly, $\hat{\theta}_j(s) = \hat{\tau}_j(s)$ for $j \leq s$ and $\hat{\theta}_j(s) = 0$ for $j > s$. Hence, it suffices to show the asymptotic covariance matrix between the nonzero random variables $\hat{\tau}_i(r)$ and $\hat{\tau}_j(s)$ with $0 \leq i \leq r$, $p_0 \leq r \leq p$, and $p+1 \leq j \leq s$, $s \leq P$, to be equal to zero. But this follows from (A4), since the asymptotic covariance matrix between $n^{1/2}(\hat{\tau}_i(r) - \tau^0_i(r))$ and $n^{1/2}(\hat{\tau}_j(s) - \tau^0_j(s))$ is found in the columns of block C^0_{uv}, $u = r - p_0 + 1$, $v = s - p_0 + 1$, with indices not less than $K(j) + 1 \geq K(p+1) + 1 \geq K(r+1) + 1 > K(r)$.

To prove the first half of the lemma observe that $S = n^{1/2}((\hat{\theta}(p_0) - \theta^0)', \ldots, (\hat{\theta}(P) - \theta^0)')'$ is obtained from T by adding components which are identically zero at the appropriate positions; hence, S is asymptotically normal with mean zero and covariance matrix D^0 which is made up of $K(P) \times K(P)$-blocks D^0_{uv} which for $u \leq v$ are given by

$$D^0_{uv} = \begin{bmatrix} (A^0_{p_0+u-1})^{-1} & 0 \\ 0 & 0 \end{bmatrix}. \tag{A5}$$ ■

Proof of Lemma 4. We have $\text{pr}(\hat{p}_n = p) = \text{pr}(\hat{t}_P \leq c_P, \ldots, \hat{t}_{p+1} \leq c_{p+1}, \hat{t}_p > c_p)$ by definition, if we adopt the convention $\hat{t}_0 = \infty$ and c_0 arbitrary but finite. Consider first the case $p > p_0$. For $p \leq r \leq P$ Assumptions A and B(i) imply that $\hat{t}_r$ differs from $t_r = (a^0_r/n)^{-1/2}|\hat{\tau}_r(r)|$ only by a term which goes to zero in probability, that the variables t_r are asymptotically independent, and that the limit distribution of t_r is the distribution of the absolute value of a standard normal random variable. Hence, $\lim_{n\to\infty} \text{pr}(\hat{t}_r \leq c_r) = 1 - \alpha_r$ by the definition of c_r. Using the asymptotic independence we arrive at $\gamma_p = \alpha_p(1 - \alpha_{p+1}) \cdots (1 - \alpha_P)$. Next consider the case $p = p_0$. Since $\text{pr}(\hat{t}_{p_0} \leq c_{p_0})$ goes to zero as $n \to \infty$ in view of Assumption B(ii) and the above convention if $p_0 = 0$, γ_{p_0} differs from $\text{pr}(\hat{t}_P \leq c_P, \ldots, \hat{t}_{p_0+1} \leq c_{p_0+1})$ only by a term which goes to zero as $n \to \infty$. Evaluating the limit of this latter expression by a similar argument as above gives $\gamma_{p_0} = (1 - \alpha_{p_0+1}) \cdots (1 - \alpha_P)$. Finally, for $p < p_0$ the result follows, since $\text{pr}(\hat{p}_n = p) \leq \text{pr}(\hat{t}_{p_0} \leq c_{p_0}) \to 0$ as $n \to \infty$ in view of Assumption B(ii). ■

Proof of Theorem 1. We start from the relation

$$\text{pr}(n^{1/2}(\tilde{\theta} - \theta^0) \leq x) = \sum_{p=0}^{P} \text{pr}(n^{1/2}(\tilde{\theta} - \theta^0) \leq x, \hat{p}_n = p). \tag{A6}$$

Denote $\text{pr}(n^{1/2}(\tilde{\theta} - \theta^0) \leq x \mid \hat{p}_n = p)$ by $F_{n,p}(x)$ and observe that $F_{n,p}$ is well-defined for $p \geq p_0$ and n large enough because of $\gamma_p > 0$. Since $\gamma_p = 0$ for $p < p_0$ was shown in Lemma 4, (A6) implies

$$\lim_{n\to\infty} \text{pr}(n^{1/2}(\tilde{\theta} - \theta^0) \leq x) = \sum_{p=p_0}^{P} \gamma_p \lim_{n\to\infty} F_{n,p}(x)$$

provided the limits on the right-hand side exist. Now for any p, $p_0 \leq p \leq P$, we have $F_{n,p}(x) = 0$ if $x_i < 0$ for some $i > p$, since $\theta^0_i = 0$ for $i > p$ and $\tilde{\theta}_i = 0$ for $i > p$ on the event $\{\hat{p}_n = p\}$. Hence, assume $x_i \geq 0$ for all $i > p$. Then

$$\lim_{n\to\infty} F_{n,p}(x) = \gamma_p^{-1} \lim_{n\to\infty} \text{pr}(n^{1/2}(\hat{\tau}(p) - \tau^0(p)) \leq x[p], \\ \hat{t}_P \leq c_P, \ldots, \hat{t}_{p+1} \leq c_{p+1}, \hat{t}_p > c_p), \tag{A7}$$

again adopting the convention for $\hat{t}_0$ and c_0 used in the preceding proof. Consider first the case $p > p_0$. In view of Lemma 3 and Assumption B(i), we may replace $\hat{t}_r$ by t_r in the last expression. Using the asymptotic independence expressed in Lemma 3 we obtain

$$\lim_{n\to\infty} F_{n,p}(x) = \gamma_p^{-1}(1 - \alpha_{p+1}) \cdots (1 - \alpha_P) \lim_{n\to\infty} \text{pr}(n^{1/2}(\hat{\tau}(p) - \tau^0(p)) \\ \leq x[p], n^{1/2}|\hat{\tau}_p(p)| > (a^0_p)^{1/2}c_p).$$

Observing that $\tau_p^0(p) = 0$ since $p > p_0$ and that $n^{1/2}(\hat{\tau}(p) - \tau^0(p))$ is asymptotically normal with mean zero and covariance matrix $(A_p^0)^{-1}$ as shown in Lemma 3, we immediately arrive at

$$\begin{aligned}\lim_{n\to\infty} F_{n,p}(x) &= \alpha_p^{-1}\{\Phi_p(x[p-1], \min(x_p, -(a_p^0)^{1/2}c_p)) \\ &\qquad + \max\{0, \Phi_p(x[p]) - \Phi_p(x[p-1], (a_p^0)^{1/2}c_p)\}\} = F_p(x).\end{aligned}$$

Now finally let $p = p_0$. Then we get from (A7)

$$\begin{aligned}&\lim_{n\to\infty} F_{n,p_0}(x) \\ &\quad = \gamma_{p_0}^{-1} \lim_{n\to\infty} \operatorname{pr}(n^{1/2}(\hat{\tau}(p_0) - \tau^0(p_0)) \le x[p_0], \hat{t}_P \le c_P, \ldots, \hat{t}_{p_0+1} \le c_{p_0+1}),\end{aligned}$$

since $\operatorname{pr}(\hat{t}_{p_0} > c_{p_0}) \to 1$ in view of Assumption B(ii) and the above convention if $p_0 = 0$. Now analogously as above we obtain

$$\lim_{n\to\infty} F_{n,p_0}(x) = \lim_{n\to\infty} \operatorname{pr}\,(n^{1/2}(\hat{\tau}(p_0) - \tau^0(p_0)) \le x[p_0]) = \Phi_{p_0}(x). \qquad \blacksquare$$

Proof of Corollary. From Theorem 1 and the discussion following this theorem we have $g_{p_0} = f_{p_0} = \phi_{p_0}$ and, in case $p > p_0$,

$$g_p(x[p_0]) = \int f_p(x[p])\,dx_{p_0+1}\cdots dx_p = \alpha_p^{-1}\int_{U(p)} \phi_p(x[p])\,dx_{p_0+1}\cdots dx_p$$

with $U(p) = \mathbb{R}^{p-p_0-1} \times U_*(p)$ and$U_*(p) = \{x_p \in \mathbb{R} : |x_p| > (a_p^0)^{1/2}c_p\}$. Factorize $\phi_p(x[p])$ into conditional and marginal densities

$$\phi_p(x[p]) = \psi_1(x_{p_0+1}, \ldots, x_{p-1}\,|\,x[p_0], x_p)\psi_2(x_p\,|\,x[p_0])\psi_3(x[p_0]).$$

Substituting this into the expression for g_p gives

$$\begin{aligned}g_p(x[p_0]) &= \psi_3(x[p_0])\alpha_p^{-1}\int_{U_*(p)}\int \psi_1\,dx_{p_0+1}\cdots dx_{p-1}\psi_2\,dx_p \\ &= \psi_3(x[p_0])\alpha_p^{-1}\int_{U_*(p)} \psi_2\,dx_p \qquad \textbf{(A8)}\end{aligned}$$

observing that the inner integral is unity since ψ_1 is a probability density. (In case $p = p_0 + 1$ the factor ψ_1 is to be omitted.) Clearly, $\psi_3(x[p_0])$ equals $\phi(x[p_0]; B_p^0)$. From standard properties of the multivariate normal distribution it follows that ψ_2 is a normal density with mean $\mu_p x[p_0]$ and variance σ_p^2. Evaluating the remaining integral in (A8) then gives

$$g_p(x[p_0]) = \phi(x[p_0]; B_p^0)\kappa_p(x[p_0], \alpha_p).$$

Furthermore, for $p > p_0$ we have

$$M_p^0 = \alpha_p^{-1}\int_{V(p)} x[p_0]x[p_0]'\phi_p(x[p])\,dx[p] \qquad \textbf{(A9)}$$

where $V(p) = \mathbb{R}^{K(p)-1} \times U_*(p)$. Now, elementary calculations show that for $p > p_0$

$$M_p^0 = B_p^0 + 2d_p^0 d_p^{0\prime}(a_p^0)^{-1}c_p\phi^*(c_p)\alpha_p^{-1} \qquad \textbf{(A10)}$$

where d_p^0 contains the first $K(p_0)$ elements of the last column of $(A_p^0)^{-1}$ and ϕ^* denotes the standard normal density. Observing that $d_p^0 d_p^{0\prime}$ is nonnegative definite and that $c_p > 0$, we arrive at $M_p^0 \geq B_p^0$. Furthermore, from (A9)

$$M_p^0 \leq \alpha_p^{-1} \int x[p_0]x[p_0]'\phi_p(x[p])\,dx[p] = \alpha_p^{-1}B_p^0.$$

Since $M_p^0 \geq B_p^0$ has just been established for $p > p_0$, since $M_{p_0}^0 = (A_{p_0}^0)^{-1} = B_{p_0}^0$ and since $B_P^0 \geq B_p^0 \geq B_{p_0}^0$ clearly holds for $P \geq p \geq p_0$, we obtain $M^0 \geq B_{p_0}^0$ from $M^0 = \sum_{p=p_0}^{P} \gamma_p M_p^0$. To establish $M^0 \leq B_P^0$ we assume $p_0 < P$, since the case $p_0 = P$ is trivial. We shall make use of the relation

$$B_{p+1}^0 = B_p^0 + (a_{p+1}^0)^{-1} d_{p+1}^0 d_{p+1}^{0\prime} \tag{A11}$$

for $p_0 \leq p < P$, which we shall prove later. Observing that $c_p\phi^*(c_p) \leq (1 - \alpha_p)/2$, using (A10), (A11) and the definition of γ_p we arrive at

$$\begin{aligned} M^0 = \sum_{p=p_0}^{P} \gamma_p M_p^0 &\leq \gamma_{p_0} B_{p_0}^0 + \gamma_{p_0+1} B_{p_0+1}^0 + \gamma_{p_0}(a_{p_0+1}^0)^{-1} d_{p_0+1}^0 d_{p_0+1}^{0\prime} \\ &+ \sum_{p=p_0+2}^{P} \{\gamma_p B_p^0 + \gamma_p(1 - \alpha_p)\alpha_p^{-1}(a_p^0)^{-1} d_p^0 d_p^{0\prime}\} = (\gamma_{p_0} + \gamma_{p_0+1}) B_{p_0+1}^0 \\ &+ \sum_{p=p_0+2}^{P} \{\gamma_p B_p^0 + \gamma_p(1 - \alpha_p)\alpha_p^{-1}(a_p^0)^{-1} d_p^0 d_p^{0\prime}\}. \end{aligned} \tag{A12}$$

If now $P = p_0 + 1$, then the last sum is empty and we have shown $M^0 \leq B_P^0$ since $\gamma_{p_0} + \gamma_{p_0+1} = 1$ in this case. If $P > p_0 + 1$, then we observe that $\gamma_{p_0} + \gamma_{p_0+1} = \gamma_{p_0+2}(1 - \alpha_{p_0+2})/\alpha_{p_0+2}$, and proceeding as before we see that the right-hand side of (A12) is not larger than

$$(\gamma_{p_0} + \gamma_{p_0+1} + \gamma_{p_0+2}) B_{p_0+2}^0 + \sum_{p=p_0+3}^{P} \{\gamma_p B_p^0 + \gamma_p(1 - \alpha_p)\alpha_p^{-1}(a_p^0)^{-1} d_p^0 d_p^{0\prime}\}. \tag{A13}$$

Proceeding now in this manner we see that M^0 is bounded by B_P^0. It remains to prove (A11). Let $\bar{B}_{p+1}$ denote the leading principal submatrix of $(A_{p+1}^0)^{-1}$ of dimension $K(p)$, let $\bar{d}_{p+1}$ be a column vector consisting of the first $K(p)$ entries of the last column of $(A_{p+1}^0)^{-1}$, and put $v = (u',0)'$ where u is $K(p_0) \times 1$ and v is $K(p) \times 1$. We then have for all u

$$\begin{aligned} u'(B_{p+1}^0 - (a_{p+1}^0)^{-1} d_{p+1}^0 d_{p+1}^{0\prime})u &= v'(\bar{B}_{p+1} - (a_{p+1}^0)^{-1} \bar{d}_{p+1} \bar{d}_{p+1}')v \\ &= v'(A_p^0)^{-1} v = u' B_p^0 u \end{aligned}$$

using the formula for the inverse of a partitioned matrix. Since the matrices under consideration are all symmetric, (A11) follows. ■