Statistica Sinica 4(1994), 29-50

A COMPARISON OF ORDER ESTIMATION PROCEDURES FOR ARMA MODELS

Benedikt M. Pötscher and Sylaja Srinivasan

Universität Wien and University of Maryland

Abstract: Order estimation procedures for autoregressive moving average models such as minimum BIC, a procedure introduced in Pötscher (1990) and variants thereof, as well as Pukkila, Koreisha & Kallinen's (1990) procedure are compared on the basis of large and small sample results. The relationship between these procedures is also discussed.

Key words and phrases: Time Series, ARMA models, order estimation, model selection, minimum BIC.

1. Introduction

We compare a number of order estimation procedures for autoregressive moving average (ARMA) models on the basis of large and small sample results. The procedures considered are minimum BIC, a procedure introduced in Pötscher (1990) and variants thereof, as well as a procedure put forward recently by Pukkila, Koreisha & Kallinen (1990), henceforth referred to as PKK (1990). PKK (1990) studied the sampling properties of their identification procedure by means of a Monte Carlo study and concluded that 'the proposed identification procedure is powerful, and can accurately identify the model structure and the order of mixed models' (PKK (1990, p.547)). They also claim that their procedure compares favourably with minimum BIC. We show that their procedure can be inconsistent. Furthermore, a simulation study shows that minimum BIC and a variant of Pötscher's (1990) procedure perform best among the procedures compared, calling PKK's (1990) conclusion into question. Section 2 discusses the various order estimation procedures and their large sample properties. The small sample properties based on a simulation study are presented in Section 3. Section 4 contains some concluding remarks.

2. Order Determination Procedures

In this section we discuss standard minimum BIC-type procedures, the PKK procedure, Pötscher's (1990) procedure and variants thereof, and present some theoretical results concerning their sampling properties.

An ARMA(p,q) model is of the form

$$a(L)y(t) = b(L)\varepsilon(t) \tag{2.1}$$

where $t \in \mathbf{Z}$ and the polynomials $a(z)$ and $b(z)$ are given by $a(z) = 1 - \Sigma_{i=1}^{p} a_i z^i$, $b(z) = 1 - \Sigma_{i=1}^{q} b_i z^i$ with $a_i \in \mathbf{R}$, $b_i \in \mathbf{R}$. The backward shift operator is denoted by L and the disturbances $\varepsilon(t)$ are (real valued) white noise, i.e., satisfy $E(\varepsilon(t)) = 0$, $E(\varepsilon(s)\varepsilon(t)) = \delta(s,t)\sigma^2$, $\sigma^2 > 0$, where $\delta(s,t)$ denotes Kronecker's delta. We make the standard stability and miniphase assumption

$$a(z) \neq 0 \quad \text{for} \quad |z| \leq 1, \tag{2.2a}$$
$$b(z) \neq 0 \quad \text{for} \quad |z| < 1. \tag{2.2b}$$

We assume throughout the paper that the data process $y(t)$ is a weakly stationary ARMA process, that is, $y(t)$ is the weakly stationary solution of an equation of the form (2.1) with a and b satisfying (2.2). Given $(y(t))$, there is then a minimal ARMA model satisfying (2.2) such that $y(t)$ is a solution of this model. We shall denote this minimal true model by

$$a^0(L)y(t) = b^0(L)\varepsilon(t). \tag{2.3}$$

The polynomials $a^0(z)$ and $b^0(z)$ do not have common factors, and their actual degrees p_0 and q_0 are called the orders of the ARMA process $(y(t))$.

The goal of any order estimation procedure is then to estimate (p_0, q_0) on the basis of n successive observations from $y(t)$.

2.1. Minimum BIC-type procedures

Minimum BIC-type procedures select the order $(\hat{p}, \hat{q})$ which minimizes

$$\psi(p,q) = \log(\hat{\sigma}_n^2(p,q)) + (p+q)C(n)/n \tag{2.4}$$

over a set $\mathcal{P}$ specified by the user. Typically, $\mathcal{P} = \{(p,q) : 0 \leq p \leq P,\ 0 \leq q \leq Q\}$ or $\{(p,q) : 0 \leq p+q \leq K\}$ where P, Q, and K are selected by the user. The penalty function $C(n)$, for example, is equal to $\log(n)$ for BIC, and is equal to 2 for AIC. The estimator $\hat{\sigma}_n^2(p,q)$ in (2.4) is the (Gaussian pseudo) maximum likelihood estimator for the residual variance when fitting ARMA(p,q) models. (For a more formal definition of $\hat{\sigma}_n^2(p,q)$ see Pötscher (1990, Section 6).) *We note that actual implementations of the various order determination procedures discussed in this paper may use approximations to $\hat{\sigma}_n^2(p,q)$ obtained from the Hannan-Rissanen algorithm or similar algorithms, or may use estimators for the residual variance different from the (Gaussian pseudo) maximum likelihood estimator.* The theoretical results will typically also hold for such approximate procedures or for procedures that use alternative residual variance estimators, although this needs to be verified on a case by case basis. Sometimes, in an effort to reduce computational burden, the set $\mathcal{P}$ is chosen to be $\{(r,r) : 0 \leq r \leq R\}$, i.e., only ARMA$(r,r)$ models are fitted, resulting in an estimator $\tilde{r}$. Of course, in this case one cannot expect to identify the true order (p_0, q_0), but only $r_0 = \max(p_0, q_0)$. Estimation of r_0 is nevertheless a useful exercise since the parameters of the true model are identified in the ARMA(r_0, r_0) specification. If estimators for (p_0, q_0) are desired, such estimators can be found in a second step starting from the ARMA$(\tilde{r}, \tilde{r})$ specification, e.g., by minimizing (2.4) over the set $\{(p,q) : p \leq \tilde{r},\ q = \tilde{r} \text{ or } p = \tilde{r},\ q \leq \tilde{r}\}$. (Cf. Remark 3.2 in Pötscher (1990) and the hybrid procedure defined below.)

Minimum BIC-type procedures require the user to specify the upper bounds P and Q (or K). (Even if P, Q or K are allowed to increase with sample size, as is the case in some of the theoretical results concerning such procedures, a particular value for these upper bounds has to be chosen in any given application of the procedure.) The bounds P, Q (or K) are not only an integral part of the definition of the procedure, but also consistency of the method can obviously hold only if P and Q satisfy $p_0 \leq P$, $q_0 \leq Q$ (or $p_0 + q_0 \leq K$ holds), which of course is an unverifiable assumption in general. Given $p_0 \leq P$, $q_0 \leq Q$ ($p_0 + q_0 \leq K$) holds, the estimator $(\hat{p}, \hat{q})$ that minimizes (2.4) over the set $0 \leq p \leq P$, $0 \leq q \leq Q$ ($0 \leq p + q \leq K$) is weakly consistent for (p_0, q_0) if $C(n)/n \to 0$ and $C(n) \to \infty$ as $n \to \infty$, and is strongly consistent if $C(n)/n \to 0$ and $\liminf_{n\to\infty}[C(n)/\log\log(n)] > 2$ as $n \to \infty$. (Apart from the condition $p_0 \leq P$, $q_0 \leq Q$ or $p_0 + q_0 \leq K$ and the conditions on the penalty term, some technical assumptions like a martingale difference assumption on the $\varepsilon(t)$'s and moment conditions are also required for the consistency result to hold. For a recent account of consistency results for minimum BIC-type procedures see Hannan (1980), Hannan & Deistler (1988, Ch.5).) The consistency result also continues to hold for any estimator $(\hat{p}, \hat{q})$ that is obtained from minimizing (2.4) over an arbitrary set $\mathcal{P}$ as long as $\mathcal{P}$ is finite and satisfies $(p_0, q_0) \in \mathcal{P}$.

2.2. Sequential procedures

Calculation of a minimum BIC-type order estimator requires fitting of all models with orders in the range $0 \leq p \leq P$, $0 \leq q \leq Q$ (or $0 \leq p + q \leq K$ or more generally with $(p,q) \in \mathcal{P}$). This will, in general, also include the estimation of a number of unidentified models. Model selection procedures which try to ameliorate some of these potentially computationally burdensome aspects of minimum BIC-type procedures have been proposed in the literature. A common feature of these methods is to examine the adequacy or goodness-of-fit of each candidate ARMA(p,q) model in a sequential manner proceeding from low order models to high order models. That is, starting with $k = 0$, at stage k all ARMA(p,q) models with $p+q=k$ are fitted to the data and the adequacy of each of these candidate models is determined according to some decision rule. If none of the models at stage k is judged to be adequate, the models at stage $k+1$ are examined. If one of the ARMA(p,q) models with $p+q=k$ passes the adequacy check, the corresponding order (p,q) is adopted as the order estimate. (If more than one model at stage k passes the adequacy check, the tie is broken by choosing the model with the smallest residual variance among the tied models.) Clearly, variants where one may only fit candidate ARMA(p,q) models with $p=q$ or with $q=0$ or, more generally, with $(p,q) \in \mathcal{M}$ for some user specified set $\mathcal{M}$, are possible. Obviously, the crucial element in any sequential procedure is the decision rule used to check adequacy of each ARMA model; and the performance of the sequential procedure will be determined by the operating characteristics of this decision rule. (It is important to keep in mind that the adequacy check in a sequential procedure should be simple. Of course, minimum BIC-type procedures (in fact any model selection procedure) can also be formally recast as sequential procedures. However, the adequacy check then will involve estimation of all models in $\mathcal{P}$ and comparison of the ψ-values of these models. Clearly, nothing is gained from such a reformulation of the procedure.)

For example, the procedure in Pötscher (1983) employing Lagrange multiplier tests for the adequacy check of candidate models follows the above sequential scheme. (This procedure is consistent if the significance levels of the Lagrange multiplier tests tend to zero at an appropriate rate as sample size increases. While this procedure has the advantage that it does not require the estimation of unidentified models (in large samples), it requires, similar to minimum BIC-type methods, however, the appropriate choice of upper bounds P and Q in order to achieve consistency. For this reason we do not consider this procedure any further.) Also, PKK's procedure as well as the procedure introduced in Pötscher (1990) have this sequential structure.

Procedures P1 and P2. The procedure in Pötscher (1990) was originally given for the case where only ARMA(r,r) models are fitted to the data, and $r_0 = \max(p_0, q_0)$ is to be estimated. It amounts to fitting ARMA(r,r) models and to selecting that order $\hat{r}$ which gives the first "local" minimum or $\psi(r,r)$. More formally, $\hat{r}$ is characterized by

$$\psi(r,r) > \psi(r+1,r+1) \quad \text{for} \quad 0 \le r < \hat{r}, \tag{2.5}$$

$$\psi(\hat{r},\hat{r}) \le \psi(\hat{r}+1,\hat{r}+1), \tag{2.6}$$

where it is understood that $\hat{r} = 0$ if (2.5) is impossible (i.e., if $\psi(0) \le \psi(1)$) and that $\hat{r} = \infty$ if (2.6) does not hold. In other words, $\hat{r}$ is the right hand end point of the interval on which ψ is strictly decreasing. Of course, $\hat{r}$ is defined without any reference to an upper bound for the order and also no such upper bound is needed for the consistency result: The estimator $\hat{r}$ is weakly consistent for r_0 if $C(n)/n \to 0$ and $C(n) \to \infty$ as $n \to \infty$, and $\hat{r}$ is strongly consistent if $C(n)/n \to 0$ and $\liminf_{n\to\infty}[C(n)/\log\log(n)] > 2$ as $n \to \infty$, (see Theorem 3.1 in Pötscher (1990)). (Again, the same technical assumptions as in the consistency result for minimum BIC-type estimators have to be assumed here.) The consistency result also shows that in large samples it is actually only necessary to estimate the models in the range $0 \leq r \leq r_0 + 1$ in order to calculate the estimator $\hat{r}$.

This procedure, called procedure P1 in the sequel, clearly can be viewed as a sequential procedure, where $\mathcal{M} = \{(p,p) : 0 \leq p < \infty\}$ and where the adequacy check consists of a comparison of $\psi(p,p)$ with $\psi(p+1,p+1)$. From a testing point of view this adequacy check amounts to a (pseudo) likelihood ratio test of the ARMA(p,p) specification against the ARMA$(p+1,p+1)$ specification at an appropriate significance level determined implicitly by the penalty function (cf., Pötscher (1990, p.168), Pötscher (1991a, Section 4, Remark 2)). (Since $\log(\hat{\sigma}_n^2(p,q))$ is only an approximation to the minimum of the negative (pseudo) log-likelihood, the test is in fact only an approximation to the (pseudo) likelihood ratio test.)

Procedure P1 can easily be generalized to the case where ARMA(p,q) models are fitted and (p_0,q_0) is to be estimated: Follow the general sequential scheme while judging a candidate ARMA(p,q) specification to be adequate if $\psi(p,q) \leq \psi(p+1,q+1)$. This generalization of procedure P1 will be called procedure P2. More formally, starting at $k = 0$, check if $\psi(p,q) \leq \psi(p+1,q+1)$ holds for some (p,q) with $p+q = k$. In this case set $(\hat{p},\hat{q})$ equal to the current (p,q). (If several pairs (p,q) with $p+q = k$ satisfy $\psi(p,q) \leq \psi(p+1,q+1)$ choose the one that has the smallest residual variance.) Otherwise set $k = k+1$ and repeat the process. Procedure P2 can be shown to be consistent for (p_0,q_0) under the same set of conditions which were used to prove consistency for procedure P1 in Theorem 3.1 in Pötscher (1990). (The proof follows the same steps as the proof of Theorem 3.1 in that reference and is therefore omitted.) In particular, $C(n)$ has to satisfy $C(n)/n \to 0$ and $C(n) \to \infty$ for weak consistency and $C(n)/n \to 0$ and $\liminf_{n\to\infty}[C(n)/\log\log(n)] > 2$ for strong consistency. Similar to procedure P1, the calculations involved in procedure P2, as well as the corresponding consistency result, do not require an upper bound on the ARMA order. Again the consistency result also shows that at most all models with $p+q \leq (p_0+1)+(q_0+1)$ will have to be estimated in large samples. Furthermore, the asymptotic properties of procedure P1 established in Pötscher (1990, Section 4), for the case where $(y(t))$ is not an ARMA process carry over to procedure P2.

We note that the consistency result for procedure P2 continues to hold for versions of P2 in which only candidate ARMA(p,q) models with $(p,q) \in \mathcal{M}$ are fitted and $\mathcal{M}$ is a user specified subset of $\{(p,q) : 0 \leq p < \infty, 0 \leq q < \infty\}$, as long as $(p_0, q_0) \in \mathcal{M}$ holds. For example, if only candidate ARMA$(p,0)$ models are fitted, i.e., $\mathcal{M} = \{(p,0) : 0 \leq p < \infty\}$, and if $(y(t))$ is an autoregressive process, i.e., $q_0 = 0$, then this version of P2 is consistent if the penalty term satisfies the conditions mentioned above. (It may be worth pointing out that in this example the adequacy check consists in comparing $\psi(p,0)$ to $\psi(p+1,1)$ and hence this version of procedure P2 is *not* equivalent to a procedure which searches for the first local minimum of $\psi(p,0)$. In fact, this is what the MPKK procedure with $m^* = 1$ defined below does in this example, and MPKK can be inconsistent. As discussed below in more detail, for the consistency of P1 and P2 it is crucial that both the AR and the MA order are increased in the adequacy check of a candidate model. Only in the special case where exclusively candidate ARMA(r,r) models are fitted, do procedures P1 and P2 reduce to a search for the first local minimum of $\psi(r,r)$.)

Hybrid Procedure. This procedure is a hybrid between P1 and a minimum BIC-type procedure. Procedure HYB is defined as follows: Determine $\hat{r}$ according to procedure P1; then estimate (p_0, q_0) by a minimizer of $\psi(p,q)$ over the set $\{(p,q) : 0 \leq p \leq \hat{r},\ q = \hat{r}, \text{ or } p = \hat{r},\ 0 \leq q \leq \hat{r}\} \cup \{(\hat{r}+1, \hat{r}),\ (\hat{r}, \hat{r}+1)\}$. From the consistency result for P1 and for minimum BIC-type procedures, it is clear that procedure HYB is consistent for (p_0, q_0) under the same set of conditions as P2 is. Similar to P2, procedure HYB does not involve an upper bound on the ARMA order. We note that from an asymptotic point of view the inclusion of $(\hat{r}+1, \hat{r})$ and $(\hat{r}, \hat{r}+1)$ into the set over which ψ is minimized in the definition of HYB is irrelevant. In finite samples, however, this inclusion provides some insurance against estimates $(\hat{r}, \hat{r})$ which are too small.

PKK Procedure. The procedure presented in PKK (1990) is — like procedure P2 — a sequential procedure but where adequacy of each candidate ARMA(p,q) model is checked as follows: Select a positive integer m^*, and fit auxiliary autoregressive models of order $0 \leq m \leq m^*$ to the residuals obtained from the ARMA(p,q) model fitted to the data $y(t)$. Calculate BIC$(m,0)$ for the auxiliary autoregressive models. (Recall BIC$(m,0) = \psi(m,0)$ with $C(n) = \log(n)$). If BIC is minimal at $m = 0$ then the ARMA(p,q) model is judged to be adequate (as the residual series appears then to be white noise) and the estimated ARMA order is (p,q); otherwise the sequential procedure is continued. (PKK (1990) do not give their tie-breaking rule. In personal communications Pukkila informed us that they broke ties by choosing the model with the smallest residual variance among the tied models.) We note that similarly as with the other procedures discussed in this paper clearly a penalty term different from $\log(n)$ could also be used in the definition of the PKK procedure.

The PKK procedure is closely related to the following modified PKK procedure (MPKK procedure) which we introduce here only to make comparison between procedures easier. For MPKK the adequacy check of each ARMA(p,q) model in the PKK procedure is replaced by the following step: For each trial ARMA(p,q) model the auxiliary nested set of ARMA models with orders (p,q), $\ldots,(p+m^*,q)$ is considered. The ARMA(p,q) model is judged to be adequate if its BIC value, BIC(p,q), is minimal among the BIC values, BIC$(p+m,q)$, $0 \le m \le m^*$, of the auxiliary ARMA$(p+m,q)$ models. Of course, the adequacy check of the ARMA(p,q) model in the MPKK procedure is equivalent to testing the ARMA(p,q) specification against all the ARMA$(p+m,q)$ specifications using (pseudo) likelihood ratio tests (the significance levels being determined by the penalty function of BIC). The ARMA(p,q) specification is rejected if at least one (pseudo) likelihood ratio test rejects. Clearly, the PKK procedure has a similar testing interpretation. (Indeed, whereas the MPKK procedure uses the (pseudo) maximum likelihood estimators for the residual variances of the auxiliary ARMA$(p+m,q)$ models, the PKK procedure can be viewed as using two-step estimators for these quantities that are obtained as the residual variances of auxiliary AR(m) models fitted to the residuals obtained from the ARMA(p,q) model.)

The PKK procedure as well as the MPKK procedure do not provide consistent estimators for the ARMA order, in general, as shown below in the following example. In particular, the choice of m^* is crucial. The example also shows that the claim on p.771 of Pukkila & Krishnaiah (1988) that their white noise test has power approaching unity as sample size increases is incorrect.

Example. Assume that the data process $y(t)$ is generated according to $y(t) = \varepsilon(t)+\alpha\varepsilon(t-2)$, with $|\alpha| \le 1$, $E(\varepsilon(t)) = 0$, $E(\varepsilon(t)^4) < \infty$, $\varepsilon(t)$ i.i.d., and let $m^* = 1$. Note that $\gamma(1) = 0$, where $\gamma(1)$ denotes the autocorrelation at lag one. Since the residuals from the "fitted" ARMA$(0,0)$ model are obviously the data $y(t)$, both the PKK and the MPKK procedure first compare BIC$(0,0)$ and BIC$(1,0)$ for the AR(0) and AR(1) processes fitted to the data $y(t)$. Now, BIC$(1,0) =$ BIC$(0,0) + n^{-1}\log(n) + \log(1-\hat{\gamma}(1)^2)$ where $\hat{\gamma}(1)$ is the sample correlation at lag one. Since $\gamma(1) = 0$ it follows that $n^{1/2}\hat{\gamma}(1)$ is asymptotically normal with mean zero, and hence $n\hat{\gamma}(1)^2$ is asymptotically distributed as a (nonnegative) multiple of a chi-square with one degree of freedom. Now, by the mean-value theorem, $\log(1-\hat{\gamma}(1)^2) = -\hat{\gamma}(1)^2/(1+\xi_n)$, where ξ_n is a mean-value satisfying $-\hat{\gamma}(1)^2 \le \xi_n \le 0$ and hence converges to zero in probability. From this we see that pr(BIC$(1,0) >$ BIC$(0,0)) =$ pr$(\log(n) - n\hat{\gamma}(1)^2/(1+\xi_n) > 0)$ tends to one. Hence the PKK as well as the MPKK procedure accept $(p,q) = (0,0)$ incorrectly as the true order even as the sample size increases to infinity. (We have used here the Yule-Walker estimators for the parameters of the AR(1) model rather than the maximum likelihood estimators for the sake of simplicity of the argument. This is of course immaterial to the conclusion drawn.) This example in fact shows that the PKK as well as the MPKK procedure with $m^* = 1$ will stop at the order $(0,0)$ with probability approaching unity whenever the data generating process has $\gamma(1) = 0$ (and is such that $n\hat{\gamma}(1)^2$ is asymptotically chi-square), and hence these procedures are inconsistent for such data generating processes (except if the process is white noise). Clearly, we also arrive at the same conclusion for any alternative penalty term satisfying $C(n) \to \infty$. (If $C(n)$ is constant (and positive) as is, e.g., the case for AIC, then the probability of stopping at $(0,0)$ will still be positive although it does no longer approach unity.) Similar examples can also be constructed for any other value of m^*.

Although the *calculations* for the PKK and MPKK order estimator do not depend on an upper bound for the ARMA order, the example clearly shows that the choice of the auxiliary bound m^* is crucial for consistency of these procedures. It transpires from the example that the values of m^* such that consistency is possible are linked to the true model and hence to the true order (p_0, q_0) of the data process. Therefore, a selection of m^* implies a bound on the true orders (p_0, q_0) for which the PKK or MPKK procedure with the given choice of m^* can possibly be a consistent procedure. It seems that there are no results available in the literature establishing conditions (in particular on m^*) under which the PKK (or MPKK) procedure is consistent. While we expect that both procedures can be shown to be consistent *provided* m^* is chosen appropriately relative to the true order (p_0, q_0), the translation of this into precise conditions on m^* is less than obvious. (We can show that the MPKK procedure does not overestimate (p_0, q_0) asymptotically regardless of the choice of m^* (m^* fixed), and we conjecture that the same is true for the PKK procedure.)

Possible Extensions. (i) Procedures P1, P2 and HYB use only one alternative model in the adequacy check of a candidate ARMA(p,q) model, while the PKK and MPKK procedure use m^* such models. Of course, one could also consider a variant of procedure P1, say, that checks adequacy of an ARMA(r,r) model by comparing $\psi(r,r)$ not only with $\psi(r+1,r+1)$ but with $\psi(r+j,r+j)$, $j \leq m^*$. The ARMA(p,q) model would be considered adequate if $\psi(r+j,r+j)$, $0 \leq j \leq m^*$, is minimized at $j = 0$. Procedure P2 could be generalized similarly, and a generalization of procedure HYB would simply use the generalized P1 procedure just described instead of the original P1 procedure in the first stage. As discussed above, procedures P1, P2 and HYB are consistent. The consistency proof also shows that the generalizations of P1, P2 and HYB share this property under exactly the same conditions and regardless of the choice of m^* (m^* fixed). While the choice of m^* is an important factor for consistency or inconsistency of procedures PKK and MPKK, it is irrelevant for the consistency of the (generalized) P1, P2 and HYB procedures. Of course, in finite samples the choice of m^* may have an effect on any of the procedures, (cf. Section 3).

(ii) We can also come up with a variant of procedure P2 which is related to P2 in a similar way that PKK is to MPKK. For this variant of procedure P2 the adequacy check for the ARMA(p, q) model consists in fitting an ARMA$(1, 1)$ model to the residuals from the ARMA(p, q) model. The ψ-values of the ARMA$(0, 0)$ and ARMA$(1, 1)$ model for the residual series are then compared. If the ψ-value at $(0, 0)$ does not exceed the value at $(1, 1)$, the residuals are considered to be white and the ARMA(p, q) model is accepted. (Of course, this procedure corresponds to $m^* = 1$. Again it could be generalized to the case $m^* > 1$.)

2.3. Theoretical comparison of procedures

Relationships Between Procedures. The MPKK procedure, and hence the PKK procedure, are more closely related to the standard minimum BIC method than might appear at first glance. For example, if only AR models are fitted as candidate models, then the MPKK procedure corresponds to minimizing BIC over moving windows $\{p, \ldots, p + m^*\}$ and to choosing the smallest p for which the minimum of BIC over the corresponding window $\{p, \ldots, p + m^*\}$ occurs at p. In particular, if $m^* = 1$, MPKK reduces to a search for the first "local" minimum of BIC in the pure AR case. If ARMA models are considered then the MPKK procedure corresponds to minimizing BIC over moving sets of auxiliary ARMA models corresponding to the set of orders $\{(p, q), (p + 1, q), \ldots, (p + m^*, q)\}$, and to estimating the order by the "first" (p, q) which is the minimizer of BIC in its corresponding auxiliary set of models.

Clearly, the MPKK (and the PKK) procedure are also close relatives of procedures P1 and P2. Instead of using the auxiliary ARMA$(p + m, q)$ models with $0 \leq m \leq m^*$ to judge the adequacy of a candidate model, procedure P2 uses only the ARMA$(p + m, q + m)$ models with $0 \leq m \leq m^* = 1$ for that purpose. (As mentioned in Section 2.2 procedures P1 and P2 could also be generalized to the case $m^* > 1$.) The important difference between the MPKK (PKK) procedure on the one hand and procedures P1 and P2 on the other hand is that in the former procedure only the AR order is increased in the adequacy check of an ARMA model, while in the adequacy check in the latter procedures both the AR and the MA order are increased simultaneously. One ramification of this difference is that procedure P2 (and also P1) is consistent (if $C(n)$ satisfies the appropriate conditions) whereas the MPKK (and PKK) procedure is not. The central reason for this different behaviour of the two kinds of procedures is the following property of $\sigma^2(p,q)$, the limit of $\hat{\sigma}_n^2(p,q)$ as $n \to \infty$: Similar to Pötscher (1990) for the case $p = q$, it can be shown that $\sigma^2(p+1,q+1) < \sigma^2(p,q)$ holds if $p < p_0$ or if $q < q_0$. The point here is that this is a strict inequality. (It is not hard to see that $\sigma^2(p',q') \leq \sigma^2(p,q)$ always holds for $p' \geq p$ and $q' \geq q$.) Hence, if $C(n)/n \to 0$ holds, procedure P2 will not stop asymptotically before (p_0, q_0) is reached as is now easily seen. (Furthermore, $\sigma^2(p,q) = \sigma^2(p_0,q_0)$ if $p \geq p_0$ and $q \geq q_0$. Based on this, procedure P2 can be shown to actually stop at (p_0, q_0) in large samples if the penalty term satisfies the conditions for the consistency result. This is proved similarly to the corresponding result for minimum BIC-type estimators in Hannan (1980), (cf. also Pötscher (1990)).) However, the strict inequality $\sigma^2(p+1,q) < \sigma^2(p,q)$ need *not* hold if $p < p_0$ or if $q < q_0$ and this suggests that in such a case the MPKK and PKK procedures (with $m^* = 1$) will stop before (p_0, q_0) is reached. The Example in Section 2.2 shows that this is indeed the case. (Also $\sigma^2(p+m,q) = \sigma^2(p,q)$ for $0 \leq m \leq m^*$ and $m^* > 1$ is possible and the MPKK and PKK procedures will be inconsistent in such a case.)

The above discussion can be reformulated in terms of (pseudo) likelihood ratio tests. Recall that the P2 and MPKK procedures test the adequacy of a given candidate ARMA(p,q) model by a (pseudo) likelihood ratio test (with critical value proportional to $C(n)$) against "artificial" alternatives, the alternatives being ARMA$(p+1,q+1)$ for the P2 procedure and ARMA$(p+m,q)$, $1 \leq m \leq m^*$, for the MPKK procedure. It follows that the error-probabilities of the first and second kind of the (pseudo) likelihood ratio test (with critical value proportional to $C(n)$) of an incorrect ARMA(p,q) model (i.e., p or q too small) against an ARMA$(p+1,q+1)$ alternative always approach zero as the sample size increases as long as $C(n) \to \infty$ and $C(n)/n \to 0$, even if the ARMA$(p+1,q+1)$ alternative itself is incorrect (i.e., $p+1$ or $q+1$ still too small). However, this is not so if only the AR order is increased for the alternative, i.e., the (pseudo) likelihood ratio test underlying the MPKK procedure does not necessarily detect misspecification of the candidate model with probability approaching one as sample size increases. (Indeed the probability of an error of the second kind can go to one in view of the fact that the critical value goes to infinity, (cf. the Example in Section 2.2). Even if the critical value were held fixed, the probability of an error of the second kind would not vanish.) The consistency of the (pseudo) likelihood ratio test based on the ARMA$(p+1,q+1)$ alternative is closely related to identifiability issues in ARMA models. (See Pötscher (1990, 1991b) for a discussion.)

Another implication of the above discussion is that procedures that search for the first local minimum of ψ may or may not be consistent, depending on the set of ARMA models considered. For example, procedure P1 searches for the first local minimum over the set of all ARMA(r,r) models and is consistent (if $C(n)$ satisfies the appropriate conditions), while MPKK (with $m^* = 1$) — in case only

AR models are considered — searches for the first local minimum over the set of all AR models, and is not consistent in general.

Upper Bounds and Consistency. On comparing minimum BIC-type procedures, procedure P2 (and P1), procedure HYB, as well as the MPKK and PKK procedures on the basis of the large sample results discussed above, it transpires that procedure P2 (and P1) as well as procedure HYB have — in contrast to the other procedures — the advantage of not making use of an upper bound on the ARMA order at all, and of allowing consistency to be established without any unverifiable upper bound condition on p_0 and q_0 like $p_0 \leq P$ and $q_0 \leq Q$, or $p_0 + q_0 \leq K$. The MPKK and the PKK procedure also seem not to involve an upper bound on the ARMA order at first glance. However, both procedures require the choice of the upper bound m^* on the order of the auxiliary autoregressive models. As shown above, both procedures are inconsistent if this auxiliary bound is set at too small a value. Since the minimal value which allows consistency to be established is related to p_0 and q_0, it is seen that the MPKK as well as the PKK procedure, like minimum BIC-type procedures, involve an unverifiable upper bound assumption concerning p_0 and q_0.

In their paper PKK (1990) do not specify how m^* was selected. In personal communication Pukkila revealed that m^* was set at $n^{\frac{1}{2}}$ and claimed that hence the procedure does not involve an upper bound on the orders, and that this was an advantage of the PKK procedure compared with minimum BIC which requires the specification of upper bounds P and Q for the ARMA orders. However, this does not seem to be a fair comparison since we can also allow the upper bounds P and Q in minimum BIC-type procedures to increase with sample size. For minimum BIC-type procedures with slowly increasing P and Q (essentially for P and Q of order not larger than $\log(n)$) consistency has been established, (see, e.g., Hanann & Deistler (1988, Ch.5), for an extensive discussion). The consistency of the PKK or MPKK procedure may also hold if m^* increases slowly, although we are not aware of any proof and it may not be easy to obtain this result for the faster rate $n^{\frac{1}{2}}$. In any case, in a finite sample situation a value of m^* has to be selected.

Computational Complexity. Minimum BIC-type procedures involve estimation of all models in $\mathcal{P}$, and hence are typically more computationally demanding than the other procedures discussed. The computations for minimum BIC-type model selection will typically also require estimation of unidentified models. Computationally, procedures P1 and P2 (with $m^* = 1$) seem to compare favorably with their competitors, since for the adequacy check in these procedures only one alternative model has to be estimated and this model will have to be estimated at a later stage anyway (except when the procedures already stop before this latter stage is reached). Also, procedure HYB has roughly the same computational complexity as P2. The computational complexity of the PKK procedure is comparable with the complexity of P2, although contrarily to P2, the computations involved in the adequacy check can not be used at a later stage in the procedure.

The procedures P1, P2 and HYB (with $m^* = 1$) require only the estimation of one unidentified model in large samples if $C(n)$ satisfies the conditions for consistency. PKK's procedure was designed to avoid the necessity for estimating unidentified models. If our conjecture in Section 2.2 that the PKK procedure is consistent *provided* m^* is chosen appropriately is true, PKK's procedure, indeed, avoids estimation of unidentified models. (We note, however, that the MPKK procedure may require the estimation of unidentified ARMA models even if m^* has been chosen appropriately.) While estimation of unidentified models may pose numerical problems (e.g., the search for the optimum may take many iterations or may fail to converge numerically), it does not pose a problem for minimum BIC-type procedures and for procedures P1, P2, and HYB from a statistical point of view. In fact, if the algorithm employed does not locate the global optimum of the likelihood function when an *unidentified* model is estimated, this may even be helpful, as then the estimate of the residual variance and hence the ψ-value will tend to be larger, making it easier to recognize the model as overparameterized. (We note that for the algorithms employed we did not encounter any numerical difficulties in estimating the residual variance or residual series of unidentified models.)

Overestimation vs. Underestimation. It is rather obvious from the definitions of the procedures P2, PKK and MPKK that they will tend to choose lower order models in small samples than the minimum BIC-type procedure (if the upper bound on the models considered by the minimum BIC-type procedure exceeds the true orders, and if the same penalty term is used in all the procedures). This observation is also confirmed by the simulation study. Hence, for a fair comparison of all these procedures it seems that one also has to consider higher order models such that the tendency of these procedures to choose lower orders and not only the proneness of minimum BIC-type procedures to overestimate can be evaluated. In their simulation study, PKK (1990) considered only low order models, hence their results give too favourable an impression of the performance of the PKK procedure. Also, PKK (1990) compare their procedure with a minimum BIC-type procedure (i.e., the procedure of Hannan & Kavalieris (1984)) only for a few low order moving average processes and when only ARMA(r, r) models are estimated, (cf. Table 5 in PKK (1990)). Given the tendency of the PKK procedure to underestimate, the choice is then effectively between $(0, 0)$ and $(1, 1)$. Since the moving average processes considered there show strong correlation, it is not surprising that $(0,0)$ is not selected frequently by the PKK procedure. (Furthermore, in that comparison, the ARMA models are estimated by different methods for the minimum BIC procedure and the PKK procedure, respectively. Hence, differences in the performances of the order selection procedures may to some extent result from the different estimators used for the ARMA parameters.)

3. Small Sample Properties

The small sample properties of minimum BIC, PKK, P2, HYB, and HYB4 were investigated by means of a simulation study. Here HYB refers to the hybrid procedure (with $m^* = 1$), while HYB4 refers to the variant of this procedure with $m^* = 4$ discussed at the end of Section 2.2. For all procedures considered the penalty term $C(n)$ was set equal to $\log(n)$. All calculations were done in RATS386, Version 3.11. Using the random number generator available in RATS386, 100 replications of an independent identically $N(0,1)$ distributed sequence $(\varepsilon(t))$ of length 150 were generated. Various ARMA(p,q) processes were generated from the sequence $(\varepsilon(t))$ by solving $a(L)y(t) = b(L)\varepsilon(t)$ for $y(t)$ (for $t \geq \max(p,q)$ and setting all presample values of $y(t)$ equal to zero, i.e., $y(t) = 0$ for $t < \max(p,q)$). The last 100 observations of the resulting series were then used as the realization of the ARMA(p,q) process in the simulation study. The same ARMA structures as in PKK (1990) were used. These structures are low order ARMA processes. To evaluate the performance of the various procedures for higher order models, a few higher order processes were also included into our study. (See the Appendix for a list of the ARMA processes used.)

Each of the model selection procedures described above involves fitting of ARMA(p,q) models and calculation of the residual series from the fitted model. For the former purpose we used the linear procedure described in Koreisha and Pukkila (1990) as do PKK (1990) in their simulation study. (The main reason for choosing this estimation procedure was to keep our results comparable with the results in PKK (1990).) The residuals from an estimated ARMA(p,q) model were obtained as follows: First the (y_t) series, $1 \leq t \leq n$, was transformed by the estimated autoregressive operator leading to a series (v_t), say, with $p+1 \leq t \leq n$. Next the Choleski decomposition of the variance covariance matrix of dimension $n-p$ of a MA(q) process with unit error variance and parameters equal to the estimated parameters in the moving average part of the ARMA(p,q) model was calculated. Finally, the inverse of the Choleski factor was multiplied with the vector $(v_{p+1},\ldots,v_n)'$ in order to obtain the residual series. The sample variance of the residuals was then used as an estimator for $\sigma^2(p,q)$. (PKK (1990) do not describe in their paper how the residuals were obtained. From personal communication with T. Pukkila we gathered that they followed the procedure just described.) The estimation procedure of Koreisha and Pukkila (1990) sometimes produces noninvertible models (especially when underfitting the model). Therefore, for the minimum BIC, P2, HYB and HYB4 procedures, whenever the estimated model was noninvertible, Wilson's (1969) factorization algorithm was used to obtain an equivalent invertible ARMA(p, q) model before the residuals were obtained by the Choleski factorization just described. (We ignore here the possibility of obtaining roots exactly on the unit circle.) For the PKK procedure it is not necessary to replace noninvertible models with the equivalent invertible ones for the following reason: In case the estimated model is noninvertible, the variance of the corresponding residuals will be proportional to the variance of the (estimated) innovations where the proportionality factor is unknown. Now, since the PKK procedure does not compare the BIC-values over different ARMA models but only compares the BIC-values of different AR models fitted to the residual series of an estimated ARMA model, the proportionality factor does not matter. The procedure of Koreisha and Pukkila (1990) used to fit ARMA models involves estimating an initial long autoregression followed by a generalized least squares estimation step. As in PKK (1990), the long autoregression was fitted by Burg's algorithm, (cf. Andersen (1974)), and its order was set equal to the square root of sample size, i.e., equal to 10. We used one iteration in the generalized least squares step.

Ties in the PKK and P2 procedures were broken by choosing the model with the smallest residual variance among the tied models. For minimum BIC we searched over all ARMA(p, q) models with $p + q \leq K$, $K = 5$ for low order processes (cf. Table 1) and $K = 10$ for high order processes (cf. Table 2). The value of m^* in the PKK procedure, i.e., the highest order of autoregressive models fitted to the residuals from the ARMA models, was set equal to the square root of the sample size i.e., equal to 10, the same value as used by PKK (1990) in their study (personal communication). As in PKK (1990), these autoregressions were estimated by the Yule-Walker method.

We want to stress that all conclusions concerning the model selection procedures drawn in the following pertain to the variants of these procedures where $\sigma^2(p, q)$ is estimated by the linear procedure described above. If $\sigma^2(p, q)$ is estimated, e.g., by maximum likelihood, this may lead to improved performance of the procedures. This point deserves further investigation.

The number of correct order selections for each of the procedures and for low order models are shown in Table 1. (In Tables 1, 2 and 3 the numbers reported in parenthesis are the numbers of correct order selections if $\hat{\sigma}_n^2(p, q)$ is alternatively calculated only from the last $100 - u$ elements in the residual series obtained from the estimated ARMA(p, q) model in order to avoid end-effects in the calculation of the residuals. The value of u is specified in the tables.) The Monte Carlo standard error for these numbers can easily be estimated by $[j(n-j)/n]^{\frac{1}{2}}$ where j is the Monte Carlo estimate for the number of correct classifications reported in the tables and n is sample size. Comparing with Table 2 in PKK (1990) we find that the percentages of correct classification of the PKK procedure are almost always lower than the values reported in PKK (1990), and frequently differ by more than one, sometimes even by more than two, estimated standard errors. (The ARMA2, 1/3 process in Table 1 is not included in Table 2 of PKK (1990), but discussed on p.544 of that paper. Results for processes MA1/4, MA1/6 and MA2/3 are given in PKK (1990), Table 5, only for sample size 200 and only for the case where ARMA(r, r) models are fitted.) These discrepancies seem to indicate that the results of PKK (1990) give too optimistic an impression of the small sample performance of the PKK procedure even for low order models. From Table 1 we also see that the sequential procedures PKK and P2 are "very good if they are good, and very bad if they are bad." The minimum BIC as well as both HYB procedures are less likely to break down than are the PKK and P2 procedure, the HYB procedures generally giving better results than the minimum BIC procedure, (cf. the results for models ARMA1, 1/1 and ARMA2, 1/3 in Table 1 for example). Also, both procedures HYB and HYB4 are almost as good as PKK and P2 in most of the cases favourable to the latter procedures. For the low order models considered, HYB and HYB4 show almost identical results except for the ARMA2, 1/3 model.

The frequency distributions of selected orders obtained from the Monte Carlo experiment also confirm the remark in Section 2.3 that the PKK and P2 procedures tend to choose smaller orders than minimum BIC. (For lack of space these frequency distributions are not presented here, but are available on request.) In particular, the frequency distribution of the orders selected by PKK (P2) is concentrated almost exclusively on orders (p, q) with $p \leq p_0$ or $q \leq q_0$, whereas minimum BIC, HYB and HYB4 also selected models with $p \geq p_0$ and $q \geq q_0$. In the majority of these cases $p = p_0, q \geq q_0$ or $p \geq p_0, q = q_0$ holds, i.e., the true structure is still identified in the chosen model.

Next we discuss some results for low order models (cf. Table 1) in more detail. The AR1/1 process shows relatively little correlation and hence all procedures not infrequently identified the process as white noise (PKK: 18%, BIC: 12%, P2: 44%, HYB: 12%, HYB4: 12%). Also the incorrect order $(0, 1)$ was chosen frequently (PKK: 31%, BIC: 33%, P2: 16%, HYB: 33%, HYB4: 33%). A similar remark applies to the MA1/1 process. The ARMA1, 1/1 process can be reasonably well approximated by an autoregressive process of low order as can be seen from calculating the corresponding transfer function; not surprisingly, all procedures select the order $(1, 0)$ with relatively high frequency (PKK: 82%, BIC: 44%, P2: 68%, HYB: 52%, HYB4: 52%). Minimum BIC also selects the order $(2,0)$ in 23% of the replications from this process. The number of correct classifications for this process varies enormously over all procedures: the worst case is the PKK procedure which selects the true order only 5% of the time, whereas the best procedures (HYB, HYB4) find the true order 33% of the time. Also the results for the ARMA2, 1/3 process show similar features. The smallest number of correct classifications, 56, occurs for procedure PKK, whereas the highest number is 93 for procedure HYB4 followed by 74 for HYB and 70 for minimum BIC.

While in both cases discussed above the worst procedure, i.e., PKK, chooses a misspecified model in most of the cases where it does not detect the true order, the situation is different for the worst procedures, i.e., HYB and HYB4, in case the data follow an MA2/1 or MA2/2 process. Here, in cases where the true order $(0,2)$ is not found by HYB or HYB4, these procedures most frequently choose the $(1,2)$ or $(0,3)$ model, which are correctly specified models although they each contain one superfluous parameter. (Also minimun BIC selects $(1,2)$ or $(0,3)$ not infrequently.) For the MA2/1 process the frequencies for selecting the true order $(0,2)$ or the orders $(1,2)$ or $(0,3)$ are 97 for PKK, 92 for minimum BIC, 99 for P2, 88 for HYB, and 86 for HYB4. The corresponding numbers for the MA2/2 process are similar. Comparing these figures with the frequencies of correctly selected orders for these processes given in Table 1, one sees that the figures in Table 1 give too pessimistic an impression about the actual performance of HYB and HYB4 for these processes. The same phenomenon also occurs for other processes: Procedures minimum BIC, HYB and HYB4 often select the orders (p_0, q_0+1) or (p_0+1, q_0) when they miss the true order (p_0, q_0), whereas PKK and P2 almost exclusively select misspecified models in such a case. (Note that the true structure is identified in the ARMA(p_0, q_0+1) and ARMA(p_0+1, q_0) model.) Hence, if one also counts it as a success if a (true) model with one excess parameter is chosen, procedures minimum BIC, HYB and HYB4 look even better relative to procedures PKK and P2. As another example consider the ARMA1, 2/1 process: The numbers of correct selection are 69 for PKK, 71 for minimum BIC, 80 for HYB, and 77 for HYB4. The frequencies for selecting the true order $(1,2)$ or one of the orders $(2,2)$ and $(1,3)$, however, are 69 for PKK, 76 for minimum BIC, 89 for HYB, and 87 for HYB4.

Another interesting case is the ARMA1, 2/2 process. Here all procedures do more or less well, except for P2, which selects the incorrect order $(1,0)$ in 69% of the cases. Inspection of the array of BIC-values shows, that — although there is a marked decrease in residual variance when comparing the $(0,1)$ with the $(1,2)$ model — frequently the decrease in residual variance from the $(1,0)$ to the $(2,1)$ model is not sufficient to offset the increase in the penalty term in the corresponding BIC-values. Hence procedure P2 stops at $(1,0)$. Inspection of the

BIC-array also shows that procedure P1 (which only estimates r_0) does not suffer from the same problem as procedure P2. This is also documented by the good performance of procedures HYB and HYB4 which are built upon procedure P1. Similar remarks apply also to the MA2/3 process.

We next turn to the results for higher order models given in Table 2. The overall conclusion is that no procedure does very well at the given sample size $n = 100$. We first discuss, in more detail, the results for the seasonal processes, i.e., for the AR4/1, AR4/2, MA4/1, MA4/2 and ARMA4, 4/1 processes. The safest bet seems to be the minimum BIC procedure, followed by procedures HYB4 and PKK. Procedures P2 and HYB break down completely. This is largely due to the choice of $m^* = 1$. Although the choice of m^* is irrelevant for these procedures from an asymptotic point of view as remarked earlier, we see that it may matter in finite samples. This is also borne out by the much better results for procedure HYB4. Due to the gaps in the autocorrelation function of the seasonal processes procedures P2 and HYB using $m^* = 1$ are not able to pick up the correlation with the given sample size 100, and hence frequently select the order $(0, 0)$. Setting $m^* = 4$ in the HYB4 procedure considerably improves the performance, cf. Table 2. Also the PKK procedure has a hard time selecting the correct order for seasonal processes, except for the AR4/2 and MA4/2 processes which exhibit strong correlation. Despite the fact that $m^* = 10$ for the PKK procedure, the performance of this procedure is inferior to the performance of minimum BIC. (Recall also that — contrary to the case of procedures P2 and HYB — an appropriate choice of m^* is relevant for the consistency of the PKK procedure.) The PKK procedure also shows a tendency towards choosing orders lower than the true one.

For the nonseasonal processes, i.e., for the ARMA4, 4/2 and ARMA4, 4/3 processes, all procedures fare poorly. (This does not come as a surprise, however, as these two processes are difficult to identify: the ARMA4, 4/2 process has an AR and a MA root each of order 4, whereas the AR and MA polynomials of the ARMA4, 4/3 process have each a dominant root while the other roots are closer to zero.) Procedures PKK, P2, HYB and HYB4 almost always decide on lower order structures, while minimum BIC also selects ARMA models with 8 or more parameters in a considerable number of cases. (Again this only confirms remarks made in Section 2.3.)

In order to get more insight into the performance of the various order estimation procedures for higher order models the Monte Carlo experiment was repeated also for sample size $n = 200$. (The various ARMA processes were obtained from the residual series $(\varepsilon(t))$ in the same manner as before except that now the length of the $(\varepsilon(t))$ series was 250. The order of the initial long autoregression was set at 14 ($\cong n^{\frac{1}{2}}$). Also, for the PKK procedure a value of $m^* = 14$ was used.) The numbers of correct classifications are shown in Table 3. Not surprisingly there is improvement in the performance of the procedures, although procedures P2 and HYB still do not work at this sample size. Considering only the seasonal models minimum BIC seems to be the most consistent overall performer, followed by HYB4 and PKK. For the nonseasonal models there is almost no improvement compared with the result for sample size 100, except for procedures HYB and HYB4 in case of the ARMA4, 4/2 process. Nevertheless the performance (31 and 32 correct classifications, respectively) is not impressive. However, if we look at the frequencies of selecting one of the orders $(4,4), (4,5)$ or $(5,4)$ things look a bit brighter. These numbers are 46% for procedure HYB and 48% for procedure HYB4. (The corresponding numbers for the other procedures are PKK 1%, BIC 4%, P2 2%.) Hence the overall winner of this contest at sample size 200 seems to be procedure HYB4.

From the frequency distributions one also gathers that in some cases minimum BIC shows considerable variability. This seems to be mainly due to the fact that a large number of alternative models are compared. The sequential structure of the other procedures considered helps in bringing down this variability. The price to be paid for this is a stronger tendency towards underestimation. However, the hybrid procedures seem to be a good compromise.

PKK (1990) emphasize that their procedure frequently selects plausible alternative structures if it fails to identify the correct order. (As discussed above, their procedure then selects a misspecified alternative structure almost exclusively.) As they note, this clearly happens most often if the true process can be closely approximated by a lower order one. Of course, in case a model selection procedure fails to select an order (p,q) with $p \geq p_0$, $q \geq q_0$ (i.e., if the procedure selects a misspecified model), it is desirable that the procedure selects at least a structure that closely approximates the true process. Such a feature of a model selection procedure is only natural and will be shared by most reasonable model selection procedures (since processes which have similar covariance structures will be difficult to distinguish by any procedure). Clearly, the relative performance of model selection procedures taking into account the quality of approximation to the true model in case a misspecified model is selected is of interest, but would require evaluation of, e.g., the prediction performance of the selected models.

4. Conclusion

From the large sample as well as small sample results discussed above procedure HYB4 as well as minimum BIC seem to emerge as the best among the model selection procedures compared. The PKK and P2 procedure, while doing well in many cases, produce disastrous results in others. We also observed smaller numbers of correct selection for the PKK procedure in our simulations than reported in PKK (1990). Furthermore, the PKK procedure is shown to be an inconsistent procedure in general. On the basis of the above evidence PKK's (1990) conclusion that the PKK procedure is superior to minimum BIC type criteria seems questionable. The Monte Carlo results also show that procedures PKK and P2 most frequently choose a misspecified model (i.e., $p < p_0$ or $q < q_0$) whenever they miss the true order (p_0, q_0), while procedures minimum BIC, HYB and HYB4 often select a true, albeit, slightly overparameterized model (typically with one superfluous parameter) in such a case. Hence if one counts it a success if a model selection procedure selects the true order (p_0, q_0), or $(p_0 + 1, q_0)$ or $(p_0, q_0 + 1)$, then the performance of minimum BIC, HYB and HYB4 relative to PKK and P2 is even better than one might think from looking at Tables 1-3.

It seems that in cases where different model selection procedures disagree on the order selected, inspecting the array of BIC(p, q)-values for interesting features — apart from calculating the various model selection procedures — is good practice. Also, examining the sensitivity of the order estimator obtained from procedure HYBm^* with respect to the value of m^* will provide useful information. The question whether the performance of any of the procedures will improve substantially if the residual variance is estimated by maximum likelihood remains to be investigated.

Appendix

The following ARMA processes were used in the simulation study.

AR1/1: $p = 1$, $q = 0$, $a_1 = 0.3$, AR1/2: $p = 1$, $q = 0$, $a_1 = 0.5$,
AR1/3: $p = 1$, $q = 0$, $a_1 = 0.9$, AR1/4: $p = 1$, $q = 0$, $a_1 = -0.9$,
ARMA1, 1/1: $p = 1$, $q = 1$, $a_1 = 0.8$, $b_1 = 0.5$,
ARMA1, 1/2: $p = 1$, $q = 1$, $a_1 = 0.8$, $b_1 = -0.7$,
ARMA1, 1/3: $p = 1$, $q = 1$, $a_1 = -0.8$, $b_1 = 0.7$,
AR2/1: $p = 2$, $q = 0$, $a_1 = 1.42$, $a_2 = -0.73$,
AR2/2: $p = 2$, $q = 0$, $a_1 = 1.8$, $a_2 = -0.9$,
MA1/1: $p = 0$, $q = 1$, $b_1 = 0.3$, MA1/2: $p = 0$, $q = 1$, $b_1 = 0.5$,
MA1/3: $p = 0$, $q = 1$, $b_1 = 0.9$, MA1/4: $p = 0$, $q = 1$, $b_1 = -0.8$,
MA1/5: $p = 0$, $q = 1$, $b_1 = -0.9$, MA1/6: $p = 0$, $q = 1$, $b_1 = -0.95$,
MA2/1: $p = 0$, $q = 2$, $b_1 = 1.42$, $b_2 = -0.73$,
MA2/2: $p = 0$, $q = 2$, $b_1 = 1.8$, $b_2 = -0.9$,
MA2/3: $p = 0$, $q = 2$, $b_1 = -1$, $b_2 = -0.89$,
ARMA2, 1/1: $p = 2$, $q = 1$, $a_1 = 1.4$, $a_2 = -0.6$, $b_1 = -0.8$,
ARMA2, 1/2: $p = 2$, $q = 1$, $a_1 = -0.5$, $a_2 = -0.9$, $b_1 = 0.6$,
ARMA2, 1/3: $p = 2$, $q = 1$, $a_1 = -0.95$, $a_2 = -0.9$, $b_1 = -0.5$,
ARMA1, 2/1: $p = 1$, $q = 2$, $a_1 = -0.8$, $b_1 = 1.4$, $b_2 = -0.6$,

ARMA1, 2/2: $p = 1$, $q = 2$, $a_1 = 0.6$, $b_1 = -0.5$, $b_2 = -0.9$,
AR4/1: $p = 4$, $q = 0$, $a_1 = 0$, $a_2 = 0$, $a_3 = 0$, $a_4 = 0.5$,
AR4/2: $p = 4$, $q = 0$, $a_1 = 0$, $a_2 = 0$, $a_3 = 0$, $a_4 = 0.9$,
MA4/1: $p = 0$, $q = 4$, $b_1 = 0$, $b_2 = 0$, $b_3 = 0$, $b_4 = 0.5$,
MA4/2: $p = 0$, $q = 4$, $b_1 = 0$, $b_2 = 0$, $b_3 = 0$, $b_4 = 0.9$,
ARMA4, 4/1: $p = 4$, $q = 4$, $a_1 = 0$, $a_2 = 0$, $a_3 = 0$, $a_4 = 0.9$, $b_1 = 0$, $b_2 = 0$, $b_3 = 0$, $b_4 = -0.9$,
ARMA4, 4/2: $a(z) = (1 - 0.9z)^4$, $b(z) = (1 + 0.9z)^4$,
ARMA4, 4/3: $a(z) = (1 - 0.9z)(1 - 0.7z)(1 - 0.5z)(1 - 0.3z)$,
$b(z) = (1 + 0.9z)(1 + 0.7z)(1 + 0.5z)(1 + 0.3z)$.

Table 1. Low Order Models: Number of correctly estimated orders
sample size: 100, number of replications: 100

	PKK	MINIMUM BIC	P2	HYB	HYB4
MA1/1	44	43 (51)	24 (29)	45 (37)	45 (37)
MA1/2	82	76 (81)	79 (82)	80 (80)	80 (80)
MA1/3	95	88 (93)	100 (100)	90 (94)	90 (94)
MA1/4	98	89 (91)	99 (98)	92 (92)	92 (92)
MA1/5	97	88 (91)	100 (99)	90 (93)	90 (93)
MA1/6	90	91 (92)	99 (99)	92 (93)	92 (93)
AR1/1	51	51 (50)	40 (36)	53 (49)	53 (49)
AR1/2	85	77 (80)	84 (84)	85 (84)	85 (84)
AR1/3	99	96 (96)	100 (100)	97 (92)	97 (92)
AR1/4	98	92 (94)	100 (99)	98 (93)	98 (93)
MA2/1	94	76 (87)	99 (99)	58 (66)	58 (66)
MA2/2	94	87 (87)	86 (84)	64 (62)	63 (62)
MA2/3	82	85 (92)	22 (24)	82 (86)	82 (86)
ARMA1, 1/1	5	18 (17)	11 (11)	33 (33)	33 (33)
ARMA1, 1/2	98	86 (90)	97 (97)	92 (91)	92 (91)
ARMA1, 1/3	93	78 (86)	94 (96)	88 (89)	88 (89)
AR2/1	100	90 (95)	100 (100)	93 (92)	92 (91)
AR2/2	100	90 (92)	99 (99)	93 (89)	93 (89)
ARMA1, 2/1	69	71 (77)	86 (90)	80 (83)	77 (82)
ARMA1, 2/2	85	82 (85)	20 (19)	91 (90)	91 (90)
ARMA2, 1/1	93	84 (89)	94 (97)	93 (90)	93 (90)
ARMA2, 1/2	71	80 (85)	52 (49)	89 (90)	89 (90)
ARMA2, 1/3	56	70 (76)	57 (55)	74 (72)	93 (95)

Figures in parentheses are number of correctly identified models using only the last $100-u$ elements in the residual series for estimation of the residual variance. Note $u = 5$ for the BIC and P2 columns and $u = 14$ for the HYB and HYB4 columns.

Table 2. High Order Models: Number of correctly estimated orders

sample size: 100, number of replications: 100

	PKK	MINIMUM BIC	P2	HYB	HYB4
AR4/1	52	67 (66)	0 (0)	0 (0)	25 (21)
AR4/2	92	88 (92)	5 (4)	4 (11)	92 (90)
MA4/1	20	52 (54)	0 (0)	0 (0)**	5 (7)**
MA4/2	75	82 (87)	0 (0)	0 (0)	75 (72)
ARMA4, 4/1	15	62 (71)	0 (0)	7 (18)*	89 (88)*
ARMA4, 4/2	0*	1 (0)	0 (0)	7 (8)	9 (8)
ARMA4, 4/3	0	1 (0)	0 (0)	1 (0)	1 (0)

Figures in parentheses are number of correctly identified models using only the last $100-u$ elements in the residual series for estimation of the residual variance. Note $u = 10$ for the BIC and P2 columns and $u = 18$ for the HYB and HYB4 columns.

*: For 1 replication the program was stopped due to non-invertibility of a matrix.

**: For 2 replications the program was stopped due to non-invertibility of a matrix.

Table 3. High Order Models: Number of correctly estimated orders

sample size: 200, number of replications: 100

	PKK	MINIMUM BIC	P2	HYB	HYB4
AR4/1	88	88 (92)	2 (1)	0 (0)	82 (75)
AR4/2	99	94 (96)	4 (4)	3 (9)	96 (96)
MA4/1	74	90 (90)	0 (0)	0 (0)	61 (60)
MA4/2	100	94 (95)	0 (0)	0 (0)	96 (97)
ARMA4, 4/1	66	97 (96)	0 (0)	7 (12)	97 (96)
ARMA4, 4/2	1	4 (2)	2 (0)	31 (28)	32 (28)
ARMA4, 4/3	0	0 (0)	0 (0)	2 (2)	2 (2)

Figures in parentheses are number of correctly identified models using only the last $100-u$ elements in the residual series for estimation of the residual variance. Note $u = 10$ for the BIC and P2 columns and $u = 18$ for the HYB and HYB4 columns.

References

Andersen, N. (1974). On the calculation of filter coefficients for maximum entropy spectral analysis. *Geophysics* **39**, 69–72.

Hannan, E. J. (1980). The estimation of the order of an ARMA process. *Ann. Statist.* **8**, 1071–1081.

Hannan, E. J. and Kavalieris, L. (1984). A method for autoregressive-moving average estimation. *Biometrika* **71**, 273–280.

Hannan, E. J. and Deistler, M. (1988). *The Statistical Theory of Linear Systems.* John Wiley, New York.

Koreisha, S. and Pukkila, T. (1990). A generalized least-squares approach for estimation of autoregressive moving-average models. *J. Time Series Anal.* **11**, 139–151.

Pötscher, B. M. (1983). Order estimation in ARMA-models by Lagrangian multiplier tests. *Ann. Statist.* **11**, 872–885. Corrections. *Ann. Statist.* **12**, 785.

Pötscher, B. M. (1990). Estimation of autoregressive moving-average order given an infinite number of models and approximation of spectral densities. *J. Time Series Anal.* **11**, 165–179.

Pötscher, B. M. (1991a). Effects of model selection on inference. Econometric Theory **7**, 163–185.

Pötscher, B. M. (1991b). Noninvertibility and pseudo-maximum likelihood estimation of misspecified ARMA models. *Econometric Theory* **7**, 435–449. Corrections. *Econometric Theory* **10**, forthcoming.

Pukkila, T. and Krishnaiah, P. R. (1988). On the use of autoregressive order determination criteria in univariate white noise tests. *IEEE Trans. ASSP* **36**, 764–774.

Pukkila, T., Koreisha, S. and Kallinen, A. (1990). The identification of ARMA models. *Biometrika* **77**, 537–548.

Wilson, G. (1969). Factorization of the covariance generating function of a pure moving average process. *SIAM J. Numer. Anal.* **6**, 1–7.

Institut für Statistik und Informatik, Universität Wien, A-1010 Wien, Austria.
Department of Economics, University of Maryland, College Park, MD 20742, U.S.A.

(Received March 1992; accepted August 1993)