Statistica Sinica **7**(1997), 285-309

INFORMATION AND PREDICTION CRITERIA FOR MODEL SELECTION IN STOCHASTIC REGRESSION AND ARMA MODELS

Tze Leung Lai and Chang Ping Lee

Stanford University

Abstract: After a brief review of several information-based and prediction-based model selection criteria, we extend Rissanen's accumulated prediction error criterion and Wei's Fisher information criterion (FIC) from linear to general stochastic regression models, which include ARMA models and nonlinear ARX models in time series analysis as special cases. Strong consistency of these model selection criteria is established under certain conditions and the FIC is also shown to be an asymptotic approximation to some Bayes procedure. The special case of ARMA models is then studied in detail, and theoretical analysis and simulation results show that the FIC compares favorably with other procedures in the literature.

Key words and phrases: Accumulated prediction error criterion, BIC, Fisher information criterion, FPE, Kullback-Leibler information.

1. Introduction

There is a large literature on the determination of the orders p and q of ARMA models

$$y_t = a_1 y_{t-1} + \cdots + a_p y_{t-p} + \epsilon_t + b_1 \epsilon_{t-1} + \cdots + b_q \epsilon_{t-q}, \tag{1.1}$$

in which the ϵ_t are unobservable random disturbances that are assumed to form a martingale difference sequence satisfying certain assumptions. The recent monograph by Choi (1992) gives a comprehensive survey and an extensive bibliography on the subject. As pointed out by Choi, representative works include Anderson's (1963) multiple hypothesis testing, Akaike's (1969, 1974) final prediction error (FPE) and information criterion (AIC), the Box-Jenkins (1976) pattern identification methods based on the autocorrelation and partial autocorrelation functions, Parzen's (1975) criterion based on the autoregression transfer function (CAT), the Bayesian information criterion (BIC) of Akaike (1977) and Schwarz (1978), the penalty function methods of Hannan and Quinn (1979), and Rissanen's (1986a,b) minimum description length and predictive least squares principles.

In the case $q = 0$, (1.1) is a special case of stochastic regression models of the form

$$y_t = \theta^T \mathbf{x}_t + \epsilon_t, \tag{1.2}$$

in which $\{\epsilon_t\}$ is a martingale difference sequence with respect to an increasing sequence of σ-fields $\mathcal{F}_t$ generated by the current and past observations and $\mathbf{x}_t$ is $\mathcal{F}_{t-1}$-measurable. The parameter vector θ belongs to $\mathbf{R}^\kappa$ with unspecified dimension κ that has to be estimated from the data. Wei (1992) recently analyzed Rissanen's predictive least squares criterion in the stochastic regression model (1.2) under the weakest assumptions to date on the regressors $\mathbf{x}_t$, and showed that predictive least squares can be approximated by the residual sum of squares plus a penalty term that involves log det $(\sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^T)$. He called this approximation the FIC (Fisher information criterion) and discussed its advantages as a model selection criterion.

In Section 2 we derive the FIC as an asymptotic approximation to a Bayes procedure for estimating κ. We also extend the FIC and the predictive least squares criterion from (1.2) to more general stochastic regression models that include nonlinear models

$$y_t = g_t(\theta) + \epsilon_t, \tag{1.3}$$

where θ is an unknown parameter and $g_t(\theta)$ is $\mathcal{F}_{t-1}$-measurable, and prove their strong consistency under certain conditions. The general stochastic regression model (1.3) includes as special cases both (1.2), with $g_t(\theta) = \theta^T \mathbf{x}_t$, and (1.1), with $\theta = (a_1, \ldots, a_p, b_1, \ldots, b_q)^T$ and $g_t(\theta) = (y_{t-1}, \ldots, y_{t-p}, \epsilon_{t-1}(\theta), \ldots, \epsilon_{t-q}(\theta))\theta$ for $t > \max(p, q)$, where $\epsilon_i(\theta) = y_i - (y_{i-1}, \ldots, y_{i-p}, \epsilon_{i-1}(\theta), \ldots, \epsilon_{i-q}(\theta))\theta$.

In Section 3 we consider the ARMA model (1.1) and discuss certain basic issues concerning order and parameter estimation. Some asymptotic results and simulation studies are presented in this connection. They shed new light on the intricacies of order estimation in ARMA models and illustrate the usefulness of the concepts of Kullback-Leibler information and prediction errors, which have also played basic roles in Akaike's (1969, 1974) seminal work on model selection methodology. Some simulation studies are presented in Section 4.

2. Prediction- and Information-Based Model Selection Criteria in General Stochastic Regression Models

Consider the linear stochastic regression model (1.2). If the parameter vector θ is known, then the minimum variance predictor of y_{n+1} given the current and past observations $\mathbf{x}_i, y_i (i \le n)$ is $\theta^T \mathbf{x}_{n+1}$, and the optimal prediction error is ϵ_{n+1}. When θ is unknown but its dimension κ is known, the least squares predictor of y_{n+1} is $\hat{\theta}_n^T \mathbf{x}_{n+1}$, where $\hat{\theta}_n$ is the least squares estimate of θ based on $\mathbf{x}_i, y_i (i \le n)$, and the prediction error is $\epsilon_{n+1} - (\hat{\theta}_n - \theta)^T \mathbf{x}_{n+1}$. Hence

$$E(y_{n+1} - \hat{\theta}_n^T \mathbf{x}_{n+1})^2 = E(\epsilon_{n+1}^2) + E\{(\hat{\theta}_n - \theta)^T \mathbf{x}_{n+1}\}^2. \tag{2.1}$$

Assume that with probability 1,

$$E(\epsilon_{n+1}^2|\mathcal{F}_n) = \sigma^2 \text{ (positive and nonrandom) for all } n, \tag{2.2}$$

$$\lim_{n\to\infty} n^{-1}\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^T = V_\kappa \text{ (nonrandom and positive definite).} \tag{2.3}$$

Then $\sqrt{n}(\hat{\theta}_n - \theta)$ converges in distribution to a normal random vector with mean 0 and covariance matrix $\sigma^2 V_\kappa^{-1}$ (cf. Theorem 3 of Lai and Wei (1982)). Hence under uniform integrability conditions, $\lim_{n\to\infty} n\,\mathrm{Cov}(\hat{\theta}_n - \theta) = \sigma^2 V_\kappa^{-1}$. When $E(\mathbf{x}_n\mathbf{x}_n) \to V_\kappa$ and $\mathbf{x}_{n+1}$ is independent of $\mathcal{F}_n$, this implies that $nE\{\mathbf{x}_{n+1}^T(\hat{\theta}_n - \theta)\}^2 \to \sigma^2\kappa$, which reduces (2.1) to

$$E(y_{n+1} - \hat{\theta}_n^T\mathbf{x}_{n+1})^2 = \sigma^2 + n^{-1}\sigma^2\kappa + o(n^{-1}). \tag{2.4}$$

Suppose that κ is also unknown, and that one has a nested family of K candidate stochastic regression models, the kth model of which is specified by (1.2) with $\theta \in \mathbf{R}^k$ and $\mathbf{x}_t$ replaced by $\mathbf{x}_{t,k} \in \mathbf{R}^k$, so that $\mathbf{x}_{t,h}$ is an $h \times 1$ subvector of $\mathbf{x}_{t,k}$ if $h \le k$. Let $\hat{\theta}_{n,k} = (\sum_{i=1}^n \mathbf{x}_{i,k}\mathbf{x}_{i,k}^T)^{-1}\sum_{i=1}^n \mathbf{x}_{i,k}y_i$. Thus, the kth model is incorrectly specified if $k < \kappa$. The idea behind Akaike's FPE criterion to estimate κ is to first replace σ^2 in (2.4) by $(n-k)^{-1}\sum_{i=1}^n(y_i - \hat{\theta}_{n,k}^T\mathbf{x}_{i,k})^2$ and then to choose the k that minimizes this modified version of the right hand side of (2.4) up to $o(n^{-1})$ terms. Specifically, since $(1-k/n)^{-1} = 1 + k/n + o(n^{-1})$, the FPE criterion minimizes over $1 \le k \le K$

$$\mathrm{FPE}(k) = (1+2k/n)\hat{\sigma}_n^2(k), \text{ where } \hat{\sigma}_n^2(k) = n^{-1}\sum_{i=1}^{n}(y_i - \hat{\theta}_{n,k}^T\mathbf{x}_{i,k})^2. \tag{2.5}$$

Let $\mathbf{x}_t = (x_{t1}, \ldots, x_{t\kappa})^T$ and let $X_n = (x_{ij})_{1\le i\le n, 1\le j\le\kappa}$. Let $c_j(X_n)$ denote the jth column vector of X_n and let $\hat{c}_j(X_n)$ denote the projection of $c_j(X_n)$ into the linear space spanned by the other column vectors of X_n. Letting $\|\mathbf{u}\|^2 = \sum_{i=1}^n u_i^2$ for $\mathbf{u} = (u_1, \ldots, u_n)^T$, define

$$s_n^2(j) = \|c_j(X_n) - \hat{c}_j(X_n)\|^2, \quad 1 \le j \le \kappa. \tag{2.6}$$

From (2.3), it follows that $s_n^2(j)/n$ converges a.s. to some positive constant. Hence by Theorem 3.1 of Wei (1992),

$$\lim_{n\to\infty} \hat{\sigma}_n^2(k) > \sigma^2 = \hat{\sigma}_n^2(\kappa) \quad \text{a.s. for} \quad k < \kappa, \tag{2.7}$$

and therefore the FPE criterion eventually chooses some $k \ge \kappa$ with probability 1. However, the FPE criterion is typically not consistent and its limiting distribution on $\{\kappa, \kappa+1, \ldots, K\}$ has been found in certain cases (cf. Sections 3.3 and 3.9 of

Choi (1992)). This difficulty with the FPE arises from the delicate (second-order) asymptotic formula for $E(y_{n+1}-\hat{\theta}_n^T\mathbf{x}_{n+1})^2$ on which it is based. In the case where $(\mathbf{x}_i, \epsilon_i)$ are i.i.d. random vectors, instead of using an asymptotic approximation to $E(y_{n+1} - \hat{\theta}_n^T\mathbf{x}_{n+1})^2$, one can estimate this mean squared prediction error by the jackknife (or other resampling methods such as the bootstrap), leading to the cross-validation method and its variants for model selection. However, for serially correlated $(\mathbf{x}_i, y_i)$, the cross-validation approach has difficulties and also tends towards overfitting (cf. Section 3.8.3 of Choi (1992)).

Instead of using estimates and/or approximations of the "final" mean squared prediction error $E(y_{n+1}-\hat{\theta}_n^T\mathbf{x}_{n+1})^2$, Rissanen (1986b) used the accumulated prediction error

$$\mathrm{APE}(k) = \sum_{i=m}^{n}(y_i - \hat{\theta}_{i-1,k}^T\mathbf{x}_{i,k})^2, \tag{2.8}$$

which he also calls "predictive least squares", abbreviated as PLS by several authors. The m in (2.8) refers to some fixed initial sample size. Since PLS is often used to abbreviate the completely different concept of "partial least squares" in the econometrics literature and since model selection is an important problem in econometrics, we propose to use the abbreviation APE as in (2.8) instead of PLS. Moreover, APE looks closer to FPE than PLS. Unlike FPE(k), which is derived as an estimate of a second-order asymptotic approximation (2.4) to the mean squared prediction error $E(y_{n+1}-\hat{\theta}_n^T\mathbf{x}_{n+1})^2$ of the next observation, APE(k) is directly defined from the data and measures the overall performance to date of using the least squares one-step-ahead predictors when one assumes that θ has dimension k. The sum in (2.8) provides averaging of the random errors ϵ_i, and (2.8) can be analyzed as in Wei (1992) under much weaker conditions than (2.3) and other assumptions needed in the asymptotic approximation (2.4) of $E(y_{n+1} - \hat{\theta}_n^T\mathbf{x}_{n+1})^2$.

In the special case of stationary autoregressive models for which $\mathbf{x}_{t,k} = (y_{t-1}, \ldots, y_{t-k})^T$, Hannan et al. (1989) showed that

$$\mathrm{APE}(k) = n\hat{\sigma}_{n,k}^2 + \sigma^2(k\log n)(1 + o(1)) \text{ a.s.} \tag{2.9}$$

for $\kappa \le k \le K$, where $\hat{\sigma}_{n,k}^2 = n^{-1}\sum_{i=1}^{n}(y_i - \hat{\theta}_{n,k}^T\mathbf{x}_{i,k})^2$ is a consistent estimate of σ^2. From (2.9), it follows that for $\kappa \le k \le K$,

$$\log(\mathrm{APE}(k)/n) = \log\hat{\sigma}_{n,k}^2 + (kn^{-1}\log n)(1 + o(1)) \text{ a.s.} \tag{2.10}$$

Except for the $o(1)$ term, the right hand side of (2.10) is BIC(k). For the linear stochastic regression model (1.2), Wei (1992) recently generalized the asymptotic formula (2.9) to the form

$$\mathrm{APE}(k) = n\hat{\sigma}_{n,k}^2 + \sigma^2(\log\ \det\sum_{i=1}^{n}\mathbf{x}_{i,k}\mathbf{x}_{i,k}^T)(1 + o(1)) \text{ a.s.}, \tag{2.11}$$

for $\kappa \leq k \leq K$, under weak conditions on the stochastic regressors $\mathbf{x}_{i,k}$. He therefore proposed the "Fisher information criterion" which replaces APE(k) by

$$\text{FIC}(k) = n\hat{\sigma}^2_{n,k} + \hat{\sigma}^2_{n,K} \log \det(\sum_{i=1}^{n} \mathbf{x}_{i,k}\mathbf{x}^T_{i,k}) \tag{2.12}$$

and which chooses the model with the smallest FIC(k). He noted that the Fisher information criterion for model selection adds to the residual sum of squares a more natural penalty term, which reflects the amount of information required to fit the model, than simply some multiple of the number of parameters in the model as in BIC or AIC.

2.1. General stochastic regression models and accumulated prediction errors

Consider the general stochastic regression model (1.3), in which θ is an unknown parameter belonging to the interior of some compact subset Θ of $\mathbf{R}^\kappa$, $\{\epsilon_n\}$ is a martingale difference sequence with respect to $\{\mathcal{F}_n\}$ such that

$$\sup_n E(|\epsilon_n|^r|\mathcal{F}_{n-1}) < \infty \quad \text{a.s. for some} \quad r > 2, \tag{2.13}$$

and $g_t(\theta)$ is $\mathcal{F}_{t-1}$-measurable. In addition to (1.1) and (1.2), another important special case of (1.3) is the nonlinear ARX model (autoregressive model with exogeneous inputs) defined recursively by

$$y_n = f_\theta(y_{n-1}, \ldots, y_{n-p}, u_{n-d}, \ldots, u_{n-d-q}) + \epsilon_n, \tag{2.14}$$

where $\{y_n\}, \{u_n\}$ and $\{\epsilon_n\}$ denote the output, input and disturbance sequences, respectively, and $d \geq 1$ represents the delay. The input u_t at time t depends on the current and past observed inputs and outputs $y_t, u_{t-1}, y_{t-1}, \ldots, y_1, u_1$, which generate the σ-field $\mathcal{F}_t$. Hence $g_t(\theta) = f_\theta(y_{t-1}, \ldots, y_{t-p}, u_{t-d}, \ldots, u_{t-d-q})$ is $\mathcal{F}_{t-1}$-mesurable. Motivated by applications to the adaptive control problem of choosing the inputs, when the system parameter θ is unknown, so that the outputs are as close to certain target values as possible in some long-run average sense, Lai and Zhu (1991) studied adaptive predictors $\hat{y}_{t+d|t}$ that replace the unknown parameter θ in the optimal d-step ahead predictor by the least squares estimate $\hat{\theta}_t$ of θ at every stage t. In particular, for $d = 1$, the 1-step ahead adaptive predictor $\hat{y}_{t+1|t}$ at stage t is given by $\hat{y}_{t+1|t} = f_{\hat{\theta}_t}(y_t, \ldots, y_{t-p+1}, u_t, \ldots, u_{t-q})$. The performance of these adaptive predictors is measured by their accumulated prediction errors $\sum_{t=1}^{n}(y_{t+d} - \hat{y}_{t+d|t})^2$. In practice, not only is the system parameter θ unknown, but the system order (p, q) is usually not known in advance.

Thus, order and parameter estimation is a fundamental problem in system identification and is also of basic interest in the related problem of adaptive prediction and control in the stochastic system (2.14).

In the general stochastic regression model (1.3), suppose that the dimension of θ is unknown and that one has a family of regression functions $\{g_{t,k}(\lambda) : k \leq \kappa, t \geq 1, \lambda \in \Theta_k\}$ in which $g_{t,k}(\lambda)$ is $\mathcal{F}_{t-1}$-measurable for every $\lambda \in \Theta_k$. Unlike the linear model (1.2) discussed above, we do not assume here that the family of regression models is nested. Thus, vectors in Θ_k need not be subvectors of those in Θ_{k+1}. For example, suppose it is known that the dimension of θ does not exceed 3, so θ has at most three components $\theta_1, \theta_2, \theta_3$. The one-parameter model involving only θ_1 is not a submodel of the two-parameter model that involves (θ_2, θ_3). Moreover, we may have three one-parameter models in this family, corresponding to θ_1, θ_2 and θ_3 respectively. Therefore, in this example, the family of candidate regression functions consists of $g_{t,1}(\theta_1), g_{t,2}(\theta_2), g_{t,3}(\theta_3), g_{t,4}(\theta_2, \theta_3)$, etc. We shall make the following assumptions on Θ_k and $g_{t,k}$:

(i) Θ_k is a compact subset of $\mathbf{R}^{d(k)}$ for $1 \leq k \leq \kappa$, where the $d(k)$ are positive integers.

(ii) There exists κ such that the true parameter θ belongs to the interior of Θ_κ.

(iii) If θ is a subvector of some $\theta^{(k)} \in \Theta_k$, then $g_{t,k}(\theta^{(k)}) = g_{t,\kappa}(\theta)$ and $\theta^{(k)}$ belongs to the interior of Θ_k.

(iv) If θ is not a subvector of any $\lambda \in \Theta_k$, then $g_{t,k}(\lambda) \neq g_{t,\kappa}(\theta)$ for all $\lambda \in \Theta_k$.

(v) If $d(k) = d(\kappa)$ but $k \neq \kappa$, then θ is not a subvector of any $\lambda \in \Theta_k$.

The term "subvector" above does not mean proper subvector, so θ is a subvector of itself. Assumption (ii) implies that the dimension of θ is $d(\kappa)$. Assumptions (v) and (iv) imply that all other models with the same or lower dimension are incorrectly specified. Hence all models except that associated with Θ_κ are either incorrectly specified or have dimension higher than $d(\kappa)$. This ensures the "identifiability" of κ. Let

$$\hat{\theta}_t^{(k)} = \arg\min_{\lambda \in \Theta_k} \sum_{i=1}^{k} (y_i - g_{i,k}(\lambda))^2 \tag{2.15}$$

be the least squares estimate (not necessarily unique) of $\theta^{(k)} \in \Theta_k$ under the hypothesis that θ is a subvector of $\theta^{(k)}$. An obvious generalization of (2.8) to the present setting is

$$\mathrm{APE}(k) = \sum_{i=m}^{n} \{y_i - g_{i,k}(\hat{\theta}_{i-1}^{(k)})\}^2, \tag{2.16}$$

where m is some fixed initial sample size. An estimate of κ is given by

$$\kappa_n = \arg\min_{1 \leq k \leq K} \mathrm{APE}(k). \tag{2.17}$$

The minimizer of APE(k) in (2.17) may not be unique. In Subsection 2.3 we show under certain regularity conditions that with probability 1, κ_n is uniquely defined (and in fact $\kappa_n = \kappa$) for all large n.

2.2. Bayesian selection procedures and Fisher information criterion

While it is straightforward to extend the APE criterion (2.8) to general stochastic regression models (1.3), extension of the Fisher information criterion (2.12) to (1.3) is much less obvious. We shall use a Bayesian approach similar to that of Schwarz (1978) who derived the BIC as an asymptotic approximation to the Bayes procedure for estimating the dimension of the natural parameter of an exponential family based on i.i.d. observations. We first extend Schwarz's argument to derive Wei's FIC in linear stochastic regression models from similar Bayesian considerations. Suppose that in the regression model (1.2) the ϵ_t are i.i.d. normal random variables with mean 0 and known variance σ^2 and that all finite-dimensional distributions of the regressor sequence $(\mathbf{x}_{i,k})_{i\geq 1}$ do not depend on θ for every $k \leq K$, assuming the nested case in which $\mathbf{x}_{i,h}$ is an $h \times 1$ subvector of $\mathbf{x}_{i,k}$ if $h \leq k$. The dimension κ of the regression parameter θ is unknown and we put a prior probability density function $p(\cdot)$ on the set $\{1, \ldots, K\}$ of possible values of κ such that $p(k) > 0$ for every $1 \leq k \leq K$. The conditional (prior) distribution of θ given $\kappa = k$ is assumed to have a continuous and everywhere positive density function $f(\cdot|k)$ with respect to Lebesgue measure on $\mathbf{R}^k$. With this specification of the distribution of the ϵ_t and the prior distribution of the parameters κ and θ, the Bayes rule (with respect to the 0-1 loss) for selecting k is to choose the k with the highest posterior probability, or equivalently, to choose the k that minimizes

$$p_n(k) := -2 \log \Big\{ p(k) \int_{\mathbf{R}^k} \exp[-\sum_{i=1}^{n} (y_i - \lambda^T \mathbf{x}_{i,k})^2 / 2\sigma^2] f(\lambda|k) d\lambda \Big\}. \tag{2.18}$$

To analyze the FIC, Wei (1992) assumed among other conditions that $\lambda_{\min}(\sum_{i=1}^{n} \mathbf{x}_{i,k} \mathbf{x}_{i,k}^T) \to \infty$ a.s. for $k = K$ (and therefore also for all $k \leq K$), where we use $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ to denote the minimum and maximum eigenvalues of a symmetric matrix A. Under this assumption, application of Laplace's method (cf. Jensen (1995)) to the integral in (2.18) yields

$$p_n(k) = -2 \log\{ p(k) f(\hat{\theta}_{n,k}|k) (2\pi\sigma^2)^{k/2} \} + \Big\{ n\hat{\sigma}_{n,k}^2 / \sigma^2 + \log \det(\sum_{i=1}^{n} \mathbf{x}_{i,k} \mathbf{x}_{i,k}^T) \Big\} + o(1). \tag{2.19}$$

Ignoring the $O(1)$ summand $-2 \log\{p(k) f(\hat{\theta}_{n,k}|k)(2\pi\sigma^2)^{k/2}\} + o(1)$ in (2.19) and replacing σ^2 in the other summand by $\hat{\sigma}_{n,K}^2$, we obtain the FIC(k) in (2.12).

To extend the FIC to the general stochastic regression models of Subsection 2.1, we shall make the additional assumption that analogous to the linear case above, there is a known K^* for which θ is a subvector of $\theta^{(K^*)} \in \Theta_{K^*}$. Modification of the preceding argument leads to the following definition of FIC in general stochastic regression models:

$$\mathrm{FIC}(k) = n\hat{\sigma}^2_{n,k} + \hat{\sigma}^2_{n,K^*} \log\ \det\Big\{ \sum_{i=1}^n (\nabla g_{i,k}(\hat{\theta}^{(k)}_n))(\nabla g_{i,k}(\hat{\theta}^{(k)}_n))^T \Big\}, \tag{2.20}$$

where $\hat{\sigma}^2_{n,k} = n^{-1} \sum_{i=1}^n [y_i - g_{i,k}(\hat{\theta}^{(k)}_n)]^2$ is an estimate of the common variance σ^2 of the i.i.d. normal ϵ_i. The main steps are outlined below.

First, analogous to (2.19), define $p_n(k) := -2\log\{\, p(k) \int_{\mathbf{R}^k} \exp[-\sum_{i=1}^n (y_i - g_{i,k}(\lambda))^2/2\sigma^2]\ f(\lambda|k) d\lambda\}$, and note that the function $S_n(\lambda) := \sum_{i=1}^n (y_i - g_{i,k}(\lambda))^2$ has minimum value $\hat{\sigma}^2_{n,k}$ at $\lambda = \hat{\theta}^{(k)}_n$. Assuming $g_{i,k}$ to be twice continuously differentiable, note that $\nabla S_n(\lambda) = -2\sum_{i=1}^n (y_i - g_{i,k}(\lambda)) \nabla g_{i,k}(\lambda)$ and

$$\nabla^2 S_n(\lambda)/2 = \sum_{i=1}^n (\nabla g_{i,k}(\lambda))(\nabla g_{i,k}(\lambda))^T - \sum_{i=1}^n (y_i - g_{i,k}(\lambda)) \nabla^2_{i,k}(\lambda), \tag{2.21}$$

where $\nabla g(\lambda)$ and $\nabla^2 g(\lambda)$ denote the gradient vector and Hessian matrix of g at λ. Therefore if $\lambda_{\min}(\nabla^2 S_n(\hat{\theta}^{(k)}_n)) \to \infty$ a.s., then we can again apply Laplace's method to show that (2.19) holds with $\nabla^2 S_n(\hat{\theta}^{(k)}_n)$ replacing $\sum_{i=1}^n \mathbf{x}_{i,k}\mathbf{x}^T_{i,k}$. In fact, for normal ϵ_i, the least squares estimate $\hat{\theta}^{(k)}_n$ is the same as the maximum likelihood estimate and $\nabla^2 S_n(\hat{\theta}^{(k)}_n)$ is the same as the observed Fisher information matrix. Hence this argument leads to replacing $\sum_{i=1}^n \mathbf{x}_{i,k}\mathbf{x}^T_{i,k}$ in (2.12) by the observed Fisher information matrix, i.e.,

$$\mathrm{FIC}_1(k) = n\hat{\sigma}^2_{n,k} + \hat{\sigma}^2_{n,K^*} \log\ \det \nabla^2 S_n(\hat{\theta}^{(k)}_n). \tag{2.22}$$

A difficulty of using (2.22) in finite samples, however, is that $\nabla^2 S_n(\hat{\theta}^{(k)}_n)$ need not be nonnegative definite in view of (2.21). A commonly used method to avoid this difficulty in regression models, as in the Gauss-Newton instead of the Newton-Raphson method of computing the minimum of $S_n(\theta)$, is to discard the second term on the right hand side of (2.21), thus approximating $\nabla^2 S_n(\hat{\theta}^{(k)}_n)/2$ by the nonnegative definite matrix $\sum_{i=1}^n (\nabla g_{i,k}(\hat{\theta}^k_n))(\nabla g_{i,k}(\hat{\theta}^{(k)}_n)^T$. Moreover, under certain regularity conditions (e.g. Klimko and Nelson (1978), p. 631; Lai (1994), Theorem 2), it follows from martingale theory that the second term on the right hand side of (2.21) is of a smaller order of magnitude than the first term, so the preceding argument indeed leads to minimizing $\mathrm{FIC}(k)$ as defined in (2.20) as an approximate Bayes procedure. Further refinements of this model selection

procedure for sample sizes usually encountered in practice will be discussed in Section 3.

In the preceding derivation of the FIC by Bayesian arguments, we started with the assumption of known σ^2. If we remove this assumption and put a prior distribution on σ^2 whose density function (with respect to Lebesgue measure on $(0,\infty)$) satisfies certain regularity conditions, then we can again use Laplace's method as before and obtain instead of FIC(k) the following variant thereof:

$$\mathrm{FIC}_2(k) = \log \hat{\sigma}^2_{n,k} + n^{-1} \log \det\Big\{ \sum_{i=1}^{n} (\nabla g_{i,k}(\hat{\theta}^{(k)}_n))(\nabla g_{i,k}(\hat{\theta}^{(k)}_n))^T \Big\}. \tag{2.23}$$

2.3. Consistency of FIC and APE criterion in stochastic regression models

In view of the assumptions (i)-(v) for the family of candidate stochastic regression models in Subsection 2.1, it is convenient to use the following notation in the sequel:

$$I = \{1 \le k \le K : \theta \text{ is a subvector of some } \theta^{(k)} \in \Theta_k\}. \tag{2.24}$$

The analysis of APE and FIC in these general stochastic regression models is much more difficult than in the linear models considered by Wei (1992). The least squares estimates (2.15) typically do not have tractable closed-form expressions and their consistency requires stronger assumptions than those of Lai and Wei (1982) in the linear case. In particular, assuming the $g_{t,k}$ to have continuous partial derivatives $\partial g_{t,k}/\partial\lambda_j, \partial^2 g_{t,k}/\partial\lambda_i\partial\lambda_j (i \ne j), \ldots, \partial^k g_{t,k}/\partial\lambda_1, \ldots, \partial\lambda_k$, Lai (1994) showed that $\hat{\theta}^{(k)}_t$ is a strongly consistent estimate of $\theta^{(k)} (k \in I)$ under the following:

Condition (A_k). For every $\lambda \ne \theta^{(k)}$, there exist $1 < p_\lambda < 2$ and an open ball $B(\lambda)$ in Θ_k, centered at λ, such that

$$\inf_{\phi \in B(\lambda)} \sum_{i=1}^{n} \{g_{i,k}(\phi) - g_{i,k}(\theta^{(k)})\}^2 \to \infty \text{ a.s.},$$

$$\max_{\substack{1 \le r \le k \\ 1 \le j_1 < \cdots < j_r \le k}} \sum_{i=1}^{n} \int_{\phi \in B(\lambda; j_1, \ldots, j_r)} [\partial^r g_{i,k}(\phi)/\partial\phi_{j_1} \cdots \partial\phi_{j_r}]^2 d\phi_{j_1} \cdots d\phi_{j_r}$$

$$+ \sum_{i=1}^{n} [g_{i,k}(\lambda) - g_{i,k}(\theta^{(k)})]^2 = O\Big(\{ \inf_{\phi \in B(\lambda)} \sum_{i=1}^{n} [g_{i,k}(\phi) - g_{i,k}(\theta^{(k)})]^2 \}^{p_\lambda} \Big) \text{ a.s.},$$

where $B(\lambda; j_1, \ldots, j_r)$ denotes the r-dimensional sphere $\{\phi \in B(\lambda) : \phi_j = \lambda_j$ for $j \notin \{j_1, \ldots, j_r\}\}$.

We next establish strong consistency of the APE-based estimate (2.17) of κ under certain assumptions, which include for $k \in I$ the following analogue of conditions (2.11)-(2.13) of Lai and Zhu (1991) in their analysis of adaptive predictors in the nonlinear ARX model (2.14) when p and q are known. Let $\|(a_{ij})_{1\le i,j\le k}\| = \max_{1\le i,j\le k}|a_{ij}|$.

Condition (B_k). There exists in Θ_k a neighborhood U_k of $\theta^{(k)}$ such that

$$\sum_{i=1}^{n} \sup_{\phi\in U_k} (\|\nabla g_{i,k}(\phi)\|^2 + \|\nabla^2 g_{i,k}(\phi)\|^2) = O(n) \ \text{ a.s.}$$

Moreover, $n^{-1}\sum_{i=1}^{n}(\nabla g_{i,k}(\theta^{(k)}))(\nabla g_{i,k}(\theta^{(k)}))^T$ converges a.s. to a positive definite matrix and

$$\sup_{\|\phi-\theta\|\le\delta} n^{-1}\sum_{i=1}^{n}\|\nabla^2 g_{i,k}(\phi) - \nabla^2 g_{i,k}(\theta)\| = O(\delta) \ \text{ a.s. as } \ n\to\infty \ \text{ and } \ \delta\to 0.$$

Theorem 1. *Assume that the martingale difference sequence* $\{\epsilon_t\}$ *satisfies* (2.13) *and* $\lim_{n\to\infty} E(\epsilon_n^2|\mathcal{F}_{n-1}) = \sigma^2 > 0$ *a.s. Suppose that for* $k \notin I$,

$$\inf_{\phi\in\Theta_k} \sum_{i=1}^{n} [g_{i,k}(\phi) - g_{i,\kappa}(\theta)]^2 / \log n \to \infty \ a.s., \tag{2.25}$$

and that (A_k) *and* (B_k) *hold for* $k \in I$. *Then* $\hat\kappa_n \to \kappa$ *a.s.*

Proof. For $k \in I$, under (A_k) and (B_k), the same arguments as those in the proof of Theorem 1 of Lai and Zhu (1991) can be used to show that

$$\sum_{i=m}^{n} \{g_{i,k}(\theta^{(k)}) - g_{i,k}(\hat\theta_{i-1}^{(k)})\}^2 \sim \sigma^2 d(k)\log n \ \text{ a.s.}, \tag{2.26}$$

recalling that $d(k)$ is the dimension of the vector $\theta^{(k)}$. Since $g_{i,k}(\theta^{(k)})$ and $g_{i,k}(\hat\theta_{i-1}^{(k)})$ are $\mathcal{F}_{i-1}$-measurable, it follows from (1.3), (2.16) and Lemma 2(iii) of Lai and Wei (1982) that with probability 1,

$$\text{APE}(k) = \sum_{i=m}^{n} \epsilon_i^2 + (1+o(1))\sum_{i=m}^{n} \{g_{i,k}(\theta^{(k)}) - g_{i,k}(\hat\theta_{i-1}^{(k)})\}^2. \tag{2.27}$$

Combining (2.26) and (2.27) yields

$$\text{APE}(k) - \sum_{i=m}^{n} \epsilon_i^2 \sim \sigma^2 d(k)\log n \ \text{ for } \ k \in I, \tag{2.28}$$

with probability 1. For $k \notin I$, it follows from (1.3), (2.16) and Lemma 2(iii) of Lai and Wei (1982) that with probability 1,

$$\Big\{\mathrm{APE}(k) - \sum_{i=m}^{n} \epsilon_i^2\Big\}/\log n = (1+o(1))\sum_{i=m}^{n}\{g_{i,\kappa}(\theta) - g_{i,k}(\hat{\theta}_{i-1}^{(k)})\}^2/\log n \to \infty, \tag{2.29}$$

by (2.25). By assumption (v) in Subsection 2.1, κ is the unique minimizer of $d(k)$ for $k \in I$. Hence from (2.28) and (2.29), the desired conclusion $P\{\hat{\kappa}_n = \kappa \text{ for all large } n\} = 1$ follows.

Under the assumptions of Theorem 1, it can be shown that

$$\sum_{i=1}^{n}[y_i - g_{i,k}(\hat{\theta}_n^{(k)})]^2 = \sum_{i=1}^{n}\epsilon_i^2 + O(\log\log n) \text{ a.s.} \tag{2.30}$$

for $k \in I$ by using arguments analogous to (2.26), (2.29) and (3.14) of Lai and Zhu (1991). Combining (2.30) with (2.28) again yields (2.9) with k replaced by $d(k)$ and with $\hat{\sigma}_{n,k}^2 = n^{-1}\sum_{i=1}^{n}[y_i - g_{i,k}(\hat{\theta}_n^{(k)})]^2$, thus extending the result of Hannan et al. (1989) to general regression models satisfying the assumptions of Theorem 1. Note that the nested family of stationary autoregressive models considered by Hannan et al. (1989) indeed satisfies the assumptions of Theorem 1. Moreover, under the assumptions of Theorem 1 (see in particular (B_k)), $\log\ \det\{\sum_{i=1}^{n}(\nabla g_{i,k}(\hat{\theta}_n^{(k)}))(\nabla g_{i,k}(\hat{\theta}_n^{(k)}))^T\} \sim d(k)\log n$ a.s., so $\mathrm{APE}(k) = \mathrm{FIC}(k) + o(\log n)$ a.s. Hence the proof of Theorem 1 also shows the following.

Corollary. *Under the assumptions of Theorem* 1, *the minimizer of* $\mathrm{FIC}(k)$ *over* $1 \le k \le K$ *converges a.s. to* κ.

3. Refinements of FIC and Estimation in ARMA Models with Unspecified Orders

In this section we study the problem of order and parameter estimation in the ARMA model (1.1), which is a special case of (1.3), and discuss in this connection certain variants and refinements of the FIC model selection procedure in (1.3), for which nonlinearities often lead to computational difficulties and asymptotic approximations are often inadequate in finite samples. Let

$$A(z) = 1 - a_1 z - \cdots - a_p z^p,\ B(z) = 1 + b_1 z + \cdots + b_q z^q. \tag{3.1}$$

We shall call $B(z)$ the "moving average polynomial" and assume that

$$A(z) \text{ and } B(z) \text{ have all zeros outside the unit circle.} \tag{3.2}$$

The (minimal) order (p_0, q_0) of the model is defined by the property that $a_{p_0} \neq 0, b_{q_0} \neq 0$ and the polynomials $A(z)$ and $B(z)$ are relatively prime (i.e., have no common zero). It is assumed that there are known upper bounds $P \geq p_0$ and $Q \geq q_0$. Thus, in the notation of Section 2, we have here a finite collection of $K = PQ+P+Q$ models with unknown parameters $\theta^{(k)} = (a_1, \ldots, a_p, b_1, \ldots, b_q)^T$ for $1 \leq k = k(p, q) \leq K$, so that $d(k) = p + q$. We shall use $\hat{\sigma}_n^2(p, q), \mathrm{FIC}(p, q)$, etc., to denote $\hat{\sigma}_n^2(k(p, q)), \mathrm{FIC}(k(p, q))$, etc.

As pointed out in Section 1, there is an extensive literature on estimation of (p_0, q_0). In particular, Hannan (1980) showed that the BIC procedure that chooses (p, q) to minimize $\log \hat{\sigma}_n^2(p, q) + (p + q)n^{-1} \log n$ over $p \leq P, q \leq Q$ is strongly consistent if $E(\epsilon_n^2 | \mathcal{F}_{n-1}) = \sigma^2 > 0$ a.s. and $\sup_n E|\epsilon_n|^r < \infty$ for some $r > 4$. To circumvent the computational complexity in the (nonlinear) least squares estimates of θ needed for the evaluation of $\hat{\sigma}_n^2(p, q)$, particularly in the overparameterized case $p > p_0$ and $q > q_0$, Hannan and Rissanen (1982) proposed a three-stage procedure for estimating p_0, q_0 and the parameter vector θ. At the first stage, an autoregressive model of order k_n is fitted, with k_n determined by the BIC, and the estimated autoregressive coefficients $\tilde{a}_n(j)$ are used to estimate ϵ_t by $\tilde{\epsilon}_t = y_t - \sum_{j=1}^{k_n} \tilde{a}_n(j) y_{t-j}$. At the second stage, least squares regression of y_t on $y_{t-1}, \ldots, y_{t-p}, \tilde{\epsilon}_{t-1}, \ldots, \tilde{\epsilon}_{t-q}$ is carried out, first along the diagonal $p = q$ $(= 1, \ldots, \max\{P, Q\})$ to give $\tilde{\sigma}_n^2(p, p) =$ (residual sum of squares)$/n$ and to give $\mathrm{BIC}(p, p)$ with minimizer $\tilde{p}$, and then for $(\tilde{p}, q)$ and $(p, \tilde{p})$ with $0 \leq q \leq Q$ and $0 \leq p \leq P$. Minimizing the $\mathrm{BIC}(\tilde{p}, q)$ and $\mathrm{BIC}(p, \tilde{p})$ thus computed gives an estimate $(\hat{p}, \hat{q})$. The third stage assumes the order $(\hat{p}, \hat{q})$ and estimates the ARMA parameters by using a one-step Gauss-Newton iteration to solve the (nonlinear) least squares normal equations, initializing at the regression estimates obtained in the second stage. This procedure, however, does not give consistent estimates of (p_0, q_0). Hannan and Kavalieris (1984) proposed two ideas, and Huang and Guo (1990) and Bhansali (1991) proposed two other ideas to modify the second stage. Huang and Guo (1990) also applied their idea to nostationary ARMAX models.

An important advantage of the Hannan-Rissanen method is that recursive algorithms are available for the linear regression calculations in the first two stages, as provided by Hannan and Rissanen (1982) and Franke (1985). Moreover, the use of the method of scoring in the third stage which consists of a one-step Gauss-Newton iteration initialized at the second-stage parameter estimate yields great computational savings over the nonlinear least squares algorithms used in Hannan (1980) and Dunsmuir and Hannan (1976). Motivated by certain statistical and computational considerations, we introduce the following modifications, labeled (A)–(D), of the Hannan-Rissanen method.

(A) For selecting the order of the approximating (high-order) autoregressive model during the first stage, we use the FIC and impose a lower bound h_n on the order of the autoregressive models to choose from. Thus, the first stage uses FIC$(k, 0)$ to choose k_n from a set of integers between h_n and H_n, as will be discussed further in Theorem 2 and the simulation studies below.

(B) In the second stage of the Hannan-Rissanen procedure, one may encounter numerical difficulties in the least squares estimate when the matrix $\sum_{i=H_n+1}^{n} \mathbf{x}_i(p,q)\mathbf{x}_i^T(p,q)$ is nearly singular or severely ill-conditioned, where $\mathbf{x}_i(p,q) = (y_{i-1}, \ldots, y_{i-p}, \tilde{\epsilon}_{i-1}, \ldots, \tilde{\epsilon}_{i-q})^T$. Note that if we replace $\mathbf{x}_i(p,q)$ by $\mathbf{x}_i^*(p,q) = (y_{i-1}, \ldots, y_{i-p}, \epsilon_{i-1}, \ldots, \epsilon_{i-q})^T$, then for $p > p_0$ and $q > q_0$, there is multicollinearity in the regressors since $y_{i-1} = a_1 y_{i-2} + \cdots + a_{p_0} y_{i-p_0-1} + \epsilon_{i-1} + \cdots + b_{q_0}\epsilon_{i-q_0-1}$, and the matrix $\sum_{i=\max(p,q)+1}^{n} \mathbf{x}_i^*(p,q)\mathbf{x}_i^{*T}(p,q)$ is singular. We therefore drop the model ARMA(p,q) from further consideration if either $\sum_{i=H_n+1}^{n} \mathbf{x}_i(p,q)\mathbf{x}_i^T(p,q)$ is not invertible or

$$\Big\{(p+q)^{-1}\mathrm{tr}[\sum_{i=H_n+1}^{n} \mathbf{x}_i(p,q)\mathbf{x}_i^T(p,q)]^{-1}\Big\}$$
$$\times\Big\{\max[n, (p+q)^{-1}\mathrm{tr}(\sum_{i=H_n+1}^{n} \mathbf{x}_i(p,q)\mathbf{x}_i^T(p,q))]\Big\}^{\delta} > 1, \tag{3.3}$$

in which we take $0 < \delta < 1$.

(C) Unlike the Hannan-Rissanen method, order selection for the ARMA model is not performed on the basis of the second-stage results. We only use the second-stage parameter estimates to initialize the one-step Gauss-Newton iteration for the third stage. To avoid numerical difficulties when the estimated moving average polynomial $\hat{B}(z)$ at the end of the second stage is unstable (i.e., has zeros inside or on the unit circle), we also perform a stability check (cf. Kucera (1979)) on the $\hat{B}(z)$ obtained in the second stage and drop those models from further consideration if their estimated moving average polynomials are unstable.

(D) The third stage of the Hannan-Rissanen approach is now used both for parameter estimation and order selection based on the FIC. For those models not deleted by (B) and (C) at the end of the second stage, we compute the third-stage parameter estimates first along the diagonal $p = q(= 1, \ldots, \max\{P, Q\})$. We also perform a stability check on the estimated moving average polynomial obtained at the third stage. We use the third-stage estimate as the parameter estimate $(\hat{a}_1, \ldots, \hat{a}_p, \hat{b}_1, \ldots, \hat{b}_p)^T$ if it passes the stability check and use the second-stage estimate otherwise. Define

$$\mathrm{FIC}(p,q) = \sum_{i=\max(P,Q)+1}^{n} e_t^2(p,q) + \hat{\sigma}_{n,H_n}^2 \log\det\Big\{\sum_{i=\max(P,Q)+1}^{n} \zeta_i(p,q)\zeta_i^T(p,q)\Big\}, \tag{3.4}$$

where $\hat{\sigma}^2_{n,H_n}$ is the same as that used in (2.23) for the FIC of the first stage,

$$e_i(p,q) = y_i - \hat{a}_1 y_{i-1} - \cdots - \hat{a}_p y_{i-p} - \hat{b}_1 e_{i-1}(p,q) - \cdots - \hat{b}_q e_{i-q}(p,q), \quad (3.5)$$

$$\begin{aligned} &\zeta_i(p,q) + \hat{b}_1 \zeta_{i-1}(p,q) + \cdots + \hat{b}_q \zeta_{i-q}(p,q) \\ &= (y_{i-1}, \ldots, y_{i-p}, e_{i-1}(p,q), \ldots, e_{i-q}(p,q))^T. \end{aligned} \quad (3.6)$$

Stability of the polynomial $1 + \hat{b}_1 z + \cdots + \hat{b}_q z^q$ ensures that the $e_i(p,q)$ defined recursively by (3.5) do not grow at an exponential rate. We first find the minimizer $\tilde{p}$ of FIC (p,p) over $1 \le p \le \max(P,Q)$ and then minimize the FIC over $(\tilde{p}+j,q)$ and $(p,\tilde{p}+j)$ with $0 \le q \le \tilde{p}+j$, $0 \le p \le \tilde{p}+j$ and $(p,q) \ne (0,0)$, for $j \in \{0,1,-1\}$. The advantages of including also $(\tilde{p} \pm 1, q)$ and $(p, \tilde{p} \pm 1)$ will be illustrated in the simulation study in Section 4. If all models on the diagonal $(1 \le) p = q (\le \max(P,Q))$ are deleted by (B) or (C), we set $\tilde{p} = 1$ and still apply the preceding search algorithm (with $\tilde{p} = 1$). In the unlikely event that all models in this search are deleted by (B) or (C), we use the first-stage result $\mathrm{AR}(k_n)$ as the selected model.

Consistency of this modification of the inconsistent Hannan-Rissanen method is established in the following.

Theorem 2. *Assume that* (3.2) *holds for the* ARMA *model* (1.1), *in which* $\{\epsilon_n\}$ *is a martingale difference sequence such that* $E(\epsilon_n^2|\mathcal{F}_{n-1}) = \sigma^2 > 0$ *and* $\sup_n E|\epsilon_n|^r < \infty$ *a.s. for some* $r > 4$. *Then for the above modification of the Hannan-Rissanen method with*

$$h_n = c_n \log n < H_n \sim (\log n)^\rho \ , \quad \text{in which } \rho > 1 \text{ and } c_n \to \infty, \quad (3.7)$$

(*see* (A) *above*),$(\hat{p},\hat{q}) \to (p_0,q_0)$ *and* $\sqrt{n}(\hat{\theta} - \theta)$ *has a limiting* $N(0, \sigma^2 V^{-1}_{p_0,q_0})$ *distribution as* $n \to \infty$, *where* $\hat{\theta} = (\hat{a}_1, \ldots, \hat{a}_p, \hat{b}_1, \ldots, \hat{b}_q)^T$,

$$\begin{aligned} V_{p,q} &= \lim_{n\to\infty} E(\mathbf{z}_n \mathbf{z}_n^T) \text{ and} \\ \mathbf{z}_n + b_1 \mathbf{z}_{n-1} + \cdots + b_q \mathbf{z}_{n-q} &= (y_{n-1}, \ldots, y_{n-p}, \epsilon_{n-1}, \ldots, \epsilon_{n-q})^T, \end{aligned} \quad (3.8)$$

in which we set $b_j = 0$ *if* $j > q_0$.

Proof. By (3.2), $y_t = \epsilon_t + \sum_{j=1}^\infty \alpha_j \epsilon_{t-j}$ with α_j tending to 0 exponentially fast. By the conditional Borel-Cantelli lemma, the assumption $E(\epsilon_n^2|\mathcal{F}_{n-1}) = \sigma^2$ a.s. for all n implies that $\epsilon_t = o(\sqrt{t})$ a.s. Hence there exists $0 < \beta < 1$ such that $(\sum_{j \ge h_n} |\alpha_j| \sup_{i \le n} |\epsilon_i|)^2 = o(n\beta^{h_n})$ a.s. Since $h_n / \log n \to \infty$, this implies that

$$\sum_{t=H_n+1}^n \Big(\sum_{j \ge h_n} |\alpha_j \epsilon_{t-j}| \Big)^2 = o(n^2 \beta^{h_n}) \to 0 \text{ a.s.} \quad (3.9)$$

For $h_n \le k \le H_n$ and $H_n < t \le n$, let $\mathbf{x}_{t,k} = (y_{t-1}, \ldots, y_{t-p})^T$ be the regressors and $\tilde{\epsilon}_{t,k} = y_t - \mathbf{x}_{t,k}^T(\sum_{i=H_n+1}^n \mathbf{x}_{i,k}\mathbf{x}_{i,k}^T)^{-1}\sum_{i=H_n+1}^n \mathbf{x}_{i,k}y_i$ be the residuals obtained by fitting an AR(k) model to the data during the first stage of the procedure. From (3.9) together with Lemma 4.2 and the proof of Theorem 2.2 of Huang and Guo (1990), it follows that

$$\max_{h_n \le k \le H_n} \sum_{t=H_n+1}^{n} (\tilde{\epsilon}_{t,k} - \epsilon_t)^2 = O(H_n^2(\log n)(\log\log n)^3) \text{ a.s.}, \tag{3.10}$$

which implies that for $p \le P$ and $q \le Q$, with probability 1,

$$\begin{aligned} &\Big\| \sum_{i=H_n+1}^{n} \mathbf{x}_i(p,q)\mathbf{x}_i^T(p,q) - \sum_{i=H_n+1}^{n} \mathbf{x}_i^*(p,q)\mathbf{x}_i^{*T}(p,q)\Big\| \\ &= O((\log n)^{2\rho+1}(\log\log n)^3) = o(n^\delta), \end{aligned} \tag{3.11}$$

where $\mathbf{x}_i(p,q)$ and $\mathbf{x}_i^*(p,q)$ are defined in (B). As noted in (B), the matrix $\sum_{i=H_n+1}^n \mathbf{x}_i^*(p,q)\mathbf{x}_i^{*T}\ (p,q)$ is singular for $p > p_0$ and $q > q_0$. On the other hand, $n^{-1}\mathrm{tr}(\sum_{i=H_n+1}^n \mathbf{x}_i^*(p,q)\mathbf{x}_i^*(p,q))$ converges a.s. to a positive constant, while $\mathrm{tr}(A^{-1}) \ge \lambda_{\max}(A^{-1}) = 1/\lambda_{\min}(A)$ for any positive definite symmetric matrix A. Hence by (3.11),

$$\begin{aligned} &P\Big\{ \sum_{i=H_n+1}^{n} \mathbf{x}_i(p,q)\mathbf{x}_i^T(p,q) \text{ is singular or (3.3) holds, for all large } n\Big\} \\ &= 1 \text{ if } p > p_0 \text{ and } q > q_0. \end{aligned} \tag{3.12}$$

As shown by Hannan (1980), for $p < p_0$ or $q < q_0$,

$$\begin{aligned} \liminf \Big\{ \inf_{c_1,\ldots,c_p,d_1,\ldots,d_q} n^{-1}\sum_{i=1}^{n} &[y_i - c_1y_{i-1} - \cdots - c_py_{i-p} - d_1\epsilon_{i-1}(\mathbf{c},\mathbf{d}) - \cdots \\ &- d_q\epsilon_{i-q}(\mathbf{c},\mathbf{d})]^2\Big\} > \sigma^2 \text{ a.s.}, \end{aligned} \tag{3.13}$$

in which $\varepsilon_t(\theta)$ is defined in Section 1, $\mathbf{c} = (c_1, \ldots, c_p)$ and $\mathbf{d} = (d_1, \ldots, d_q)$. Let $p^* = \max(p_0, q_0)$. For $(p,q) = (p^*, q)$ or (p, p^*) with $q_0 \le q \le p^*$ and $p_0 \le p \le p^*$, defining $a_i = 0$ if $i > p_0$ and $b_j = 0$ if $j > q_0$, the polynomials $z^pA(z^{-1})$ and $z^qB(z^{-1})$ are still relatively prime, and therefore the matrix $V_{p,q}$ defined in (3.8) is positive definite (cf. Hannan (1973)) and

$$\lim_{n\to\infty} n^{-1}\sum_{i=1}^{n} \zeta_i(p,q)\zeta_i^T(p,q) = V_{p,q} \text{ a.s.}, \tag{3.14}$$

$$\sum_{i=1}^{n} e_i^2(p,q) = n\sigma^2 + O(\log\log n) \text{ a.s.}, \tag{3.15}$$

where $e_i(p,q)$ and $\zeta_i(p,q)$ are defined in (3.5) and (3.6), as can be shown by standard arguments and noting from (3.10) that the second-stage estimates converge a.s. to $(a_1,\ldots,a_p,b_1,\ldots,b_q)^T$ at the rate $o(n^{-\frac{1}{2}+\epsilon})$, for every $\epsilon>0$. It follows from (3.14) that $\log\det(\sum_1^n \zeta_i(p,q)\zeta_i^T(p,q)) \sim (p+q)\log n$ a.s. Moreover, by (3.10), the $\hat{\sigma}^2_{n,H_n}$ in (3.4) converges a.s. to σ^2. Therefore it follows from (3.12)-(3.15) that $P\{(\hat{p},\hat{q})=(p_0,q_0) \text{ for all large } n\}=1$, and the limiting normal distribution of $\sqrt{n}(\hat{\theta}-\theta)$ then follows by standard arguments.

We next discuss certain insights that the preceding study of order and parameter estimation in ARMA models provides for corresponding problems in more general stochastic regression models of the type considered in Section 2. First, computation of the least squares estimates in nonlinear regression models becomes very difficult when the dimension of the parameter is large, since one has to search for the minimizer of the sum of squares function over a large parameter space. This difficulty is further magnified by the need to perform such minimization tasks for each of a large number of competing models. It is therefore highly desirable to have a good but simple preliminary estimate so that one can use it to initialize the search for the minimizer of the sum of squares function and thereby limit the number of iterations to some manageable number. In the case of ARMA models, a key idea underlying the Hannan-Rissanen procedure is to use the second-stage estimate as a preliminary estimate and to refine it with a one-step Gauss-Newton iteration. The second-stage estimate is *linear* least squares for which the unobservable ϵ_t in the actual regressor is replaced by the residuals $\tilde{\epsilon}_t$ obtained by performing *linear* regression in a high-order approximating autoregressive model, noting that a stationary ARMA model can be expressed as an AR(∞) model. Another useful idea to reduce computational burden is to devise a systematic scheme to choose models adaptively for evaluation of the FIC (or BIC, or other criterion used), instead of directly computing the criterion for all candidate models and performing a complete search for the model that minimizes the criterion. For ARMA models, this idea is used in the Hannan-Rissanen method and modifications thereof. Strictly speaking, such computational shortcuts lead to procedures that are not the same as those involving exact least squares estimates and complete search for the model with the minimum FIC (or other criterion). However, as Theorem 2 shows for the ARMA case, they have the same large-sample statistical properties as their computationally prohibitive counterparts.

As pointed out in (B) above, overfitting may lead to computational difficulties because the matrix $C_n^{(k)} := \sum_{i=1}^n (\nabla g_{i,k}(\hat{\theta}_n^{(k)}))\,(\nabla g_{i,k}(\hat{\theta}_n^{(k)}))^T$ may become singular or severely ill-conditioned. This also poses difficulties for the FIC because the penalty term in FIC(k) is a multiple of $\log\det(C_n^{(k)})$(= sum of log-eigenvalues

of $C_n^{(k)}$). Since a sufficiently small eigenvalue of $C_n^{(k)}$ has a large negative logarithm, the FIC may even give negative penalty to an overfitted model that has severely ill-conditioned $C_n^{(k)}$. To address this shortcoming of the FIC, we can modify it by removing from further consideration models whose $C_n^{(k)}$ are singular or severely ill-conditioned, as in (B) above.

4. Simulation Studies and Discussion

We first report a simulation study on the performance of the preceding modification of the Hannan-Rissanen procedure (abbreviated by MHR in the sequel), and of three different modifications of the Hannan-Rissanen procedure given by Hannan and Kavalieris (1984), Huang and Guo (1990) and Bhansali (1991). The simulation study considers the following six ARMA models:

(I) $y_t = 0.5y_{t-1} + \epsilon_t + 0.8\epsilon_{t-1}$,

(II) $y_t = -0.64y_{t-1} - 0.7y_{t-2} + \epsilon_t + 0.8\epsilon_{t-1}$,

(III) $y_t = -0.2y_{t-1} + 0.05y_{t-2} + 0.01y_{t-3} + \epsilon_t - 0.7\epsilon_{t-1}$,

(IV) $y_t = 0.33y_{t-1} + 0.16y_{t-2} + \epsilon_t + 0.39\epsilon_{t-1} + 0.28\epsilon_{t-2} + 0.11\epsilon_{t-3}$,

(V) $y_t = 1.05y_{t-1} - 0.25y_{t-2} + \epsilon_t - 0.1\epsilon_{t-1} + 0.05\epsilon_{t-2}$,

(VI) $y_t = -0.7y_{t-1} + \epsilon_t - 1.1\epsilon_{t-1} + 0.3\epsilon_{t-2}$,

in which $\epsilon_1, \epsilon_2, \ldots$ are i.i.d. $N(0,1)$ random variables and the initial conditions are given by $y_0 = y_{-1} = \cdots = \epsilon_0 = \epsilon_{-1} = \cdots = 0$. It is assumed that upper bounds $P = 4$ and $Q = 4$ are known *a priori*. Taking $n = 500$, 100 independent realizations of $(y_1, \ldots, y_n)$ were generated. To each data set $(y_1, \ldots, y_n)$, we applied five different order and parameter estimation procedures, labelled MHR, HK, B, and HuG. The MHR procedure is the same as the preceding modified version, with modifications (A)–(D), of the Hannan-Rissanen procedure. To fit a high-order autoregressive model during the first stage of the procedure, we picked a reasonable range of 11 to 22 parameters to be estimated from $n = 500$ observations (so $\log n = 6.2$), i.e., we chose $h_n = 11$ and $H_n = 22$ in (A). In addition, we took $\delta = 0.6$ in (3.3). The results are reported in Tables 1–3. We have also tried several other values of h_n (ranging from 7 to 11), H_n (ranging from 20 to 30) and δ (ranging from 0.5 to 0.7), and have obtained similar results.

The HK procedure represents the following modification of the Hannan-Rissanen procedure by Hannan and Kavalieris (1984) (cf. the procedure ($\tilde{p}^{(1)}$, $\tilde{q}^{(1)}$) on p. 275 there). The first stage uses the AIC for selecting the order of an approximating autoregressive model. Moreover, order selection is performed using the BIC after repeating the second stage with the residuals $\tilde{\epsilon}_t$ obtained from the first stage replaced by those obtained from the second stage. The B procedure represents the modification of the Hannan-Rissanen procedure by Bhansali

(1991), who used the FPE criterion to select the order of an approximating autoregressive model during the first stage and estimated the order at the second stage by minimizing the "corrected final prediction error"

$$\mathrm{FPEC}(p,q)=\tilde{\sigma}_n^2(p,q)\Big\{1+n^{-1}k_n(p+q)[1+\max_{p'\le P,q'\le Q}\sum_{j=1}^{q'}\hat{b}_j^2(p',q')]\Big\},$$

where k_n is the order of the AR(k_n) model chosen at the first stage and $\hat{b}_j(p',q')$ represents the least squares estimate of b_j obtained at the second stage for the ARMA (p',q') model (cf. Sections 1 and 5 of Bhansali (1991)). The HuG procedure is the following modification of the Hannan-Rissanen procedure by Huang and Guo (1990). The first stage fixes a relatively high order $H(=H_n)=[c(\log n)^\alpha]$, with $\alpha>1$ and $c>0$, for the approximating autoregressive model instead of choosing the order by BIC. Moreover, instead of the ($\mathcal{F}_n$-measurable) residuals $\tilde{\epsilon}_t$ obtained in the first stage, the second stage uses the $\mathcal{F}_t$-measurable estimate $\epsilon_t^*=y_t-\psi_{t,H}^T(\sum_{i=H+1}^t\psi_{i,H}\psi_{i,H})^{-1}\sum_{i=H+1}^t y_i\psi_{i,H}$ of $\epsilon_t, n\ge t\ge H+m$, where $\psi_{i,H}=(y_{i-1},\ldots,y_{i-H})^T$. Furthermore, instead of the BIC, the order selection criterion at the second stage is based on the penalty function

$$\mathrm{CIC}(p,q)=n\tilde{\sigma}_n^2(p,q)+(p+q)c_n,$$

cf. Huang and Guo (1990), p. 1735. The procedure terminates with the second stage and uses the second-stage estimates to estimate the parameters of the selected model, instead of going to the third stage. Here, for $n=500$, we took $H=10$, $c_n=H\log n$ and $m=90$.

The results of the simulation study are given in Tables 1-3. Each table presents for two of the models the frequency distributions of the selected orders of the four procedures in 100 simulations. In addition, it gives the simulated value of the final prediction error

$$\begin{aligned}\hat{E}_{n+1}&:=E(y_{n+1}-\hat{y}_{n+1})^2\\&=\sigma^2+E(\hat{y}_{n+1}-a_1y_n-\cdots-a_{p_0}y_{n-p_0}-b_1\epsilon_{n-1}-\cdots-b_{q_0}\epsilon_{n-q})^2,\end{aligned}\quad(4.1)$$

where $\sigma^2=1$ and $\hat{y}_{n+1}$ is the adaptive predictor of y_{n+1} that replaces the unknown parameters in the minimum variance predictor by their estimates based on $y_1,\ldots,y_n$. To assess the accuracy of a parameter estimate $\hat{\theta}_{\hat{p},\hat{q}}$, we use the expected Kullback-Leibler information number KL:$=EI(\theta_{p_0,q_0},\hat{\theta}_{\hat{p},\hat{q}})$, where $\theta_{p_0,q_0}=(a_1,\ldots,a_{p_0},b_1,\ldots,b_{q_0})^T$ and for $\lambda=(a_1',\ldots,a_p',b_1',\ldots,b_q')^T$,

$$I(\theta_{p_0,q_0},\lambda)=\lim_{n\to\infty}E\{\log[f(\epsilon_n)/f(y_n-a_1'y_{n-1}-\cdots-a_p'y_{n-p}-b_1'\epsilon_{n-1}-\cdots-b_q'\epsilon_{n-q})]\},\quad(4.2)$$

in which f denotes the common density function of the ϵ_n (which are standard normal). Note that if the ARMA(p_0, q_0) model has good lower-order ARMA approximations, then the fitted ARMA $(\hat{p}, \hat{q})$ model may still be very close to the actual ARMA(p_0, q_0) model despite a substantial Euclidean distance between $\hat{\theta}_{\hat{p},\hat{q}}$ and the true parameter. A natural "distance" between estimated and true parameters of ARMA models and of more general stochastic dynamical systems is the Kullback-Leibler divergence of the estimated parameter from the true one, instead of the Euclidean distance.

The ARMA models (I) and (II) are the same as those in Bhansali's (1991) simulation experiments 1 and 2, which also showed 100% correct order selection rate for the B procedure but which showed the "corrected BIC" criterion of Hannan and Kavalieris to have correct order selection rates of 82% and 75% for the two models. Note that Hannan and Kavalieris (1984) proposed two modifications of the Hannan-Rissanen method, one of which was the corrected BIC criterion and the other of which was used in our simulation study as the HK procedure. In Table 1, both the MHR and HK procedures show correct order selection rates of over 95% and have $\hat{E}_{501} - \sigma^2$ and KL values comparable to those of the B procedure. The HuG procedure also shows 100% correct order selection rate for the ARMA models (I) and (II). However, it does not use the third stage of the Hannan-Rissanen procedure to refine the second-stage estimates and this resulted in its substantially larger $\hat{E}_{501}$ and KL values than those of the other procedures.

Table 1. Frequencies of selected orders in ARMA(1,1) model (I) and ARMA (2,1) model (II) for the MHR, HK, B and HuG procedures. Also given are the simulated values of $\hat{E}_{501}$ in (4.1) and KL:= $EI(\theta_{p_0,q_0}, \hat{\theta}_{\hat{p},\hat{q}})$, and their standard errors in parentheses.

	ARMA(1,1)				ARMA(2,1)			
(p, q)	MHR	HK	B	HuG	MHR	HK	B	HuG
(1,0)	0	0	0	0	0	0	0	0
(1,1)	97	98	100	100	0	0	0	0
(1,2)	0	0	0	0	0	0	0	0
(2,0)	0	0	0	0	0	0	0	0
(2,1)	3	1	0	0	97	96	100	100
(2,2)	0	0	0	0	3	1	0	0
(2,3)	0	0	0	0	0	1	0	0
(3,1)	0	1	0	0	0	0	0	0
(3,2)	0	0	0	0	0	1	0	0
(4,4)	0	0	0	0	0	1	0	0
$\hat{E}_{501} - \sigma^2$	.0042	.0040	.0032	.0220	.0059	.0061	.0059	.0162
	(.0011)	(.0010)	(.0006)	(.0032)	(.0009)	(.0009)	(.0010)	(.0028)
KL	.0029	.0028	.0026	.0043	.0031	.0033	.0028	.0048
	(.0003)	(.0003)	(.0003)	(.0004)	(.0004)	(.0004)	(.0003)	(.0005)

Table 2. Frequencies of selected orders in ARMA(3,1) model (III) and ARMA(2,3) model (IV) for the MHR, HK, B and HuG procedures. Also given are $\hat{E}_{501}$ and KL, and their standard errors in parentheses.

	ARMA(3,1)				ARMA(2,3)			
(p,q)	MHR	HK	B	HuG	MHR	HK	B	HuG
(0,1)	0	1	11	88	0	0	0	0
(0,2)	1	0	0	0	0	0	0	0
(1,0)	0	0	0	8	22	30	58	100
(1,1)	89	97	89	4	0	0	0	0
(1,2)	7	1	0	0	77	66	41	0
(2,1)	3	1	0	0	0	0	0	0
(2,2)	0	0	0	0	1	4	1	0
$\hat{E}_{501}-\sigma^2$	.0040	.0041	.0083	.0737	.0128	.0134	.0226	.0287
	(.0006)	(.0007)	(.0024)	(.0096)	(.0027)	(.0026)	(.0034)	(.0035)
KL	.0031	.0032	.0062	.0357	.0066	.0077	.0113	.0167
	(.0003)	(.0004)	(.0011)	(.0018)	(.0006)	(.0006)	(.0006)	(.0001)

Table 3. Frequencies of selected orders in ARMA(2,2) model (V) and ARMA (1,2) model (VI) for the MHR, HK, B and HuG procedures. Also given are $\hat{E}_{501}$ and KL, and their standard errors in parentheses.

	ARMA(2,2)				ARMA(1,2)			
(p,q)	MHR	HK	B	HuG	MHR	HK	B	HuG
(1,0)	18	23	48	100	0	0	0	0
(1,1)	3	68	51	0	0	6	98	100
(1,2)	15	7	0	0	74	42	0	0
(1,3)	0	0	0	0	0	0	0	0
(2,0)	64	0	1	0	0	0	0	0
(2,1)	0	2	0	0	26	47	2	0
(2,3)	0	0	0	0	0	1	0	0
(3,1)	0	0	0	0	0	1	0	0
(3,2)	0	0	0	0	0	2	0	0
(3,3)	0	0	0	0	0	1	0	0
$\hat{E}_{501}-\sigma^2$	.0086	.0112	.0168	.0233	.0067	.0343	.7079	.0954
	(.0017)	(.0017)	(.0030)	(.0034)	(.0011)	(.0186)	(.2165)	(.0152)
KL	.0056	.0075	.0100	.0153	.0069	.0117	.1515	.0311
	(.0005)	(.0005)	(.0005)	(.0001)	(.0007)	(.0013)	(.0445)	(.0005)

In the ARMA$(3,1)$ model of Table 2, no procedure chose the true order $(3,1)$ and the MHR, HK and B procedures each chose the lower order $(1,1)$ in over 85% of the 100 simulations. In spite of this, all three procedures gave small values of $\hat{E}_{501}-\sigma^2$ and KL. This can be explained by the fact that $y_t=-0.2y_{t-1}+\epsilon_t-0.7\epsilon_{t-1}$ is a very good approximation to the ARMA$(3,1)$

model (III). The HuG procedure chose the order $(1,1)$ only 4% of the time and concentrated on the even lower order $(0,1)$, but its $\hat{E}_{501}-\sigma^2$ and KL values were over 10 times of those of the MHR procedure. The choice $c_n = H\log n$ in the penalty term of CIC(p,q), therefore, appears to be too large for this model. Theorem 2.3 of Huang and Guo (1990) actually requires c_n to be considerably larger than $H\log n$, which makes undermodeling even more severe and further deteriorates the performance. Although the penalty function method of Huang and Guo (1990) gives the correct order as $n\to\infty$, there are practical difficulties in choosing c_n even for n as large as 500. The results of Table 2 on the ARMA$(2,3)$ model (IV) are somewhat surprising. No procedure chose the correct order $(2,3)$, and the HuG procedure chose only $(1,0)$ while the other three procedures chose mostly from $\{(1,0),(1,2)\}$. The zeros of $A(z)=1-0.33z-0.16z^2$ are -3.736 and 1.673, and those of $B(z)=1+0.39z+0.28z^2+0.11z^3$ are -2.552 and $0.003\pm 1.887\sqrt{-1}$, and therefore the zeros of $A(z)$ differ substantially from those of $B(z)$. In spite of this, there are many good lower-order ARMA approximations to the model (IV). This is reflected in Table 2 by the small KL numbers for the MHR and HK procedures.

The presence of good lower-order approximations to the ARMA(p_0,q_0) model (1.1) implies that the positive definite matrix $W(p_0,q_0)$ is nearly singular, where

$$W(p,q)=E_\pi\{(y_p,\dots,y_1,\epsilon_q,\dots,\epsilon_1)^T(y_p,\dots,y_1,\epsilon_q,\dots,\epsilon_1)\},\tag{4.3}$$

and π denotes the stationary distribution of $(y_0,\dots,y_{-p_0+1},\epsilon_0,\dots,\epsilon_{-q_0+1})^T$ under the true ARMA(p_0,q_0) model. For the ARMA$(2,3)$ model (IV), the $p+q$ eigenvalues of $W(p,q)$ are listed below:

$(p,q)=(2,3)$: 0.0005, 0.2543, 1, 1.5509, 4.7596.
$(p,q)=(3,3)$: 0.0001, 0.0139, 0.4778, 1.4429, 1.8695, 6.0437.
$(p,q)=(2,2)$: 0.0120, 0.4745, 1.5450, 4.5339.
$(p,q)=(2,1)$: 0.0124, 1.4137, 4.1392.
$(p,q)=(1,2)$: 0.2522, 1, 3.0305.

Note that $n\lambda_{\min}(W(2,3))$ and even $n\lambda_{\min}(W(2,2)), n\lambda_{\min}(W(2,1))$ are quite small for $n=500$. This means that most samples of size 500 do not contain much information to estimate all the parameters of the ARMA$(2,3)$ model even if $\epsilon_1,\dots,\epsilon_{500}$ should have been observable in addition to $y_1,\dots,y_{500}$. For these samples, it is indeed better to choose lower orders than $(2,3)$. Since the Hannan-Rissanen-type procedures in Table 2 first search the order $(\tilde{p},\tilde{p})$ along the diagonal and then among $(\tilde{p},q)$ and $(p,\tilde{p})$, it is not surprising to find in Table 2 a clump of selected orders at $(1,2)$ (corresponding to $\tilde{p}=2$) and another clump at $(1,0)$ (corresponding to $\tilde{p}=1$). Note that this phenomenon due to small values of $n\lambda_{\min}(W(2,3))$ would disappear with much larger values of n, say $n\ge 10^6$ (for which $n\lambda_{\min}(W(2,3))\ge 500$), and the MHR procedure should eventually

select the true order $(2, 3)$ with probability 1 by Theorem 2. On the other hand, the results of Table 2 show that the MHR procedure can still predict y_{501} well and fit an ARMA model with small Kullback-Leibler divergence from the true model even though it does not have enough information to pick the true order at $n = 500$.

The results of Table 3 demonstrate the advantages of including $(\tilde{p} \pm 1, q)$ and $(p, \tilde{p} \pm 1)$ besides $(\tilde{p}, q)$ and $(p, \tilde{p})$ in the MHR procedure. In particular, for the ARMA$(2, 2)$ model (V), each procedure ended up with $\tilde{p} = 1$ in over 90% of the 100 simulations. Although there is inadequate information to estimate all the parameters of this ARMA$(2, 2)$ model when $n = 500$, which resulted in the choice $\tilde{p} = 1$ for many samples, there is enough information to estimate the parameters of an approximating ARMA$(2, 0)$ model, as can be seen from the eigenvalues of the matrices (4.3) for the model (V) listed below:

$(p, q) = (2, 2)$: 0.0028, 0.6225, 1.4920, 6.3993.
$(p, q) = (2, 0)$: 0.5633, 5.9533.
$(p, q) = (1, 1)$: 0.6209, 3.6375.

By considering $\tilde{p} + 1$ in addition to $\tilde{p}$, the MHR procedure was able to select $(2, 0)$ for 60% of the time and to yield considerably smaller $\hat{E}_{501}$ and KL values than those of the other three procedures. The ARMA$(1, 2)$ model (VI) in Table 3 is the same as that in Bhansali's (1991) simulation experiment 11, which also showed the B procedure to miss completely the true order $(1, 2)$ and to choose the order $(1, 1)$ most often. Although the MHR procedure again picked $\tilde{p} = 1$ many times, it was still able to choose the true order $(1, 2)$ at these times because $\tilde{p} + 1$ was also considered besides $\tilde{p}$. The relatively large values of $\hat{E}_{501}$ and KL for the B procedure were perhaps due to its default option that "an errant root (of the estimated polynomials $\hat{A}(z)$ and $\hat{B}(z)$) was replaced by its reciprocal" (Bhansali (1991), p. 91), since such errant roots occurred quite often in fitting an ARMA$(1, 1)$ model and since replacing the errant roots by their reciprocals might have resulted in a highly inaccurate estimated transfer function.

Our second simulation study was motivated by the recent work of Pötscher and Srinivasan (1994) who modified the Hannan-Rissanen method by incorporating ideas from a different approach due to Pukkila, Koreisha and Kallinen (1990), abbreviated by PKK in the sequel. Instead of model selection via the BIC or other similar criterion, the PKK procedure proceeds as follows: For $\ell = 0, 1, \ldots,$ and $p+q = \ell (p \geq 0, q \geq 0)$, first fit the ARMA$(p, q)$ model to the data and then fit an auxiliary AR(m) model with $0 \leq m \leq m^*$ (some prescribed number) to the residuals. Note that AR(0) = ARMA$(0, 0)$ is the white-noise model $y_t = \epsilon_t$ for which there are no estimated parameters. If BIC(m) for the auxiliary autoregressive models is minimized at $m = 0$, then the PKK procedure accepts the white-noise model for the residuals of the fitted ARMA(p, q) model, stops incrementing ℓ

and chooses the ARMA(p,q) model. The fitting of ARMA(p,q) models uses a three-stage procedure whose first two stages are similar to those of Hannan and Rissanen (1982) and whose third stage uses generalized least squares instead of a one-step Gauss-Newton iteration. This is computationally much more laborious than the different methods in the preceding simulation study. Because of this we chose $n = 100$ in our second simulation study, as was also chosen in most of the simulation experiments reported by Pötscher and Srinivasan (1994) and by PKK. We also took $m^* = \sqrt{n} = 10$ in the PKK procedures, following Pötscher and Srinivasan.

In addition to the MHR and PKK procedures, our second simulation study also considers the following procedure (which will be denoted by PS) proposed by Pötscher and Srinivasan (1994). PS first chooses the ARMA$(\tilde{p},\tilde{p})$ model such that BIC$(r,r) >$ BIC$(r+1,r+1)$ for $0 \le r < \tilde{p}$ and BIC$(\tilde{p},\tilde{p}) \le$ BIC$(\tilde{p}+1,\tilde{p}+1)$, and then minimizes BIC$(p,q)$ over the set $\{(\tilde{p},q) : 0 \le q \le \tilde{p}\} \cup \{(p,\tilde{p}) : 0 \le p \le \tilde{p}\} \cup \{(\tilde{p}+1,\tilde{p}),(\tilde{p},\tilde{p}+1)\}$. The computation of BIC uses the same three-stage scheme to fit ARMA(p,q) models as in PKK. Because the PKK and PS procedures both include the white-noise model ARMA$(0,0)$ as a candidate model, we also included it for the MHR procedure in this simulation study, defining FIC$(0,0) = \sum_{i=1}^{n} y_i^2$. We chose $H_n = 10, h_n = 5$ and $\delta = 0.6$ for the MHR procedure (see (A) and (B) in Section 3). The simulation study considers models (I)–(VI) introduced at the beginning of this section together with the following AR and MA models:

(VII) $y_t = 0.5y_{t-4} + \epsilon_t$,

(VIII) $y_t = 0.1y_{t-2} - 0.5y_{t-4} + \epsilon_t$,

(IX) $y_t = \epsilon_t + 0.5\epsilon_{t-4}$.

Models (VII) and (IX) were also considered in the simulation study of Pötscher and Srinivasan (1994) whose Table 2 reports that PS completely misses the true order for these models in 100 simulation runs.

Besides the number of correctly estimated orders (in 100 simulated samples of size $n = 100$) on which the simulation studies of PKK and Pötscher-Srinivasan focused exclusively, we also considered the Kullback-Leibler information number KL and the final prediction error $\hat{E}_{101}$ defined in (4.1), as in Tables 1–3 for our first simulation study. The results show that MHR compares favorably with PS and PKK and performs much better than the other two procedures in the ARMA model (II) and in the AR models (VII) and (VIII). Moreover, our simulations showed MHR to be at least ten times faster than the other two procedures which require much greater computational effort. The fact that all three procedures completely miss the correct order for models (III) and (IV) is consistent with the results of our first simulation study reported in Table 2 and is due to the presence of good lower-order approximations to the true model.

Table 4. Kullback-Leibler information number KL, prediction error $\hat{E}_{101}$, and frequency #(C) of correctly estimated orders in 100 simulated samples from ARMA(p, q) models for the MHR, PS and PKK procedures. Standard errors are given in parentheses.

	KL			$\hat{E}_{101} - \sigma^2$			#(C)		
Model	MHR	PS	PKK	MHR	PS	PKK	MHR	PS	PKK
1,1 (I)	.0297	.0221	.0321	.0331	.0207	.0273	77	94	90
	(.0035)	(.0042)	(.0061)	(.0066)	(.0044)	(.0059)			
2,1 (II)	.0320	.1280	.0822	.0668	.2390	.0907	78	80	75
	(.0054)	(.0249)	(.0130)	(.0123)	(.0685)	(.0207)			
3,1 (III)	.0261	.0330	.0407	.0407	.0439	.0475	0	0	0
	(.0022)	(.0023)	(.0024)	(.0070)	(.0085)	(.0085)			
2,3 (IV)	.0277	.0249	.0233	.0443	.0420	.0456	0	0	0
	(.0025)	(.0014)	(.0011)	(.0068)	(.0061)	(.0062)			
2,2 (V)	.0258	.0216	.0210	.0410	.0331	.0324	0	0	0
	(.0048)	(.0015)	(.0015)	(.0070)	(.0056)	(.0055)			
1,2 (VI)	.0379	.0449	.0546	.0321	.0324	.0488	39	45	17
	(.0033)	(.0033)	(.0027)	(.0056)	(.0044)	(.0077)			
4,0 (VII)	.0779	.1823	.1081	.0772	.3551	.1829	64	0	51
	(.0100)	(.0043)	(.0107)	(.0164)	(.0478)	(.0361)			
4,0 (VIII)	.0535	.1736	.0988	.0635	.3617	.1489	69	0	50
	(.0069)	(.0013)	(.0101)	(.0103)	(.0532)	(.0309)			
0,4 (IX)	.0954	.1305	.1162	.1874	.2741	.2334	17	0	12
	(.0062)	(.0021)	(.0055)	(.0301)	(.0336)	(.0332)			

Acknowledgement

This research is supported by the National Science Foundation.

References

Akaike, H. (1969). Fitting autoregressive models for prediction. *Ann. Inst. Statist. Math.* **21**, 243-247.

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

Akaike, H. (1977). An objective use of Bayesian models. *Ann. Inst. Statist. Math.* **29**, 9-20.

Anderson, T. W. (1963). Determination of the order of dependence in normally distributed time series. In *Time Series Analysis* (Edited by M. Rosenblatt), 425-446. Wiley, New York.

Bhansali, R. J. (1991). Consistent recursive estimation of the order of an autoregressive moving average process. *Internat. Statist. Rev.* **59**, 81-96.

Box, G. E. P. and Jenkins, G. M. (1976). *Time Series Analysis, Forecasting and Control* (Revised Edition). Holden-Day, San Francisco.

Choi, B. S. (1992). *ARMA Model Identification*. Springer-Verlag, New York.

Dunsmuir, W. and Hannan, E. J. (1976). Vector linear time series models. *Adv. Appl. Probab.* **8**, 339-364.

Franke, J. (1985). A Levinson-Durbin recursion for autoregressive-moving average processes. *Biometrika* **72**, 573-581.

Hannan, E. J. (1973). The asymptotic theory of linear time-series models. *J. Appl. Probab.* **10**, 130-145.

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., McDougall, A. J. and Poskitt, D. S. (1989). Recursive estimation of autoregressions. *J. Roy. Statist. Soc. Ser. B* **51**, 217-233.

Hannan, E. J. and Quinn, B. G. (1979). The determination of the order of an autoregression. *J. Roy. Statist. Soc. Ser. B* **41**, 190-195.

Hannan, E. J. and Rissanen, J. (1982). Recursive estimation of mixed autoregressive-moving average order. *Biometrika* **69**, 81-94.

Huang, D. and Guo, L. (1990). Estimation of nonstationary ARMAX models based on the Hannan-Rissanen method. *Ann. Statist.* **18**, 1729-1756.

Jensen, J. L. (1995). *Saddlepoint Approximations*. Oxford University Press.

Klimko, L. A. and Nelson, P. I. (1978). On conditional least squares estimation for stochastic processes. *Ann. Statist.* **6**, 629-642.

Kucera, V. (1979). *Discrete Linear Control: The Polynomial Equation Approach*. John Wiley, New York.

Lai, T. L. (1994). Asymptotic properties of nonlinear least squares estimates in stochastic regression models. *Ann. Statist.* **22**, 1917-1930.

Lai, T. L. and Wei, C. Z. (1982). Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. *Ann. Statist.* **10**, 154-166.

Lai, T. L. and Zhu, G. (1991). Adaptive prediction in non-linear autoregressive models and control systems. *Statist. Sinica* **1**, 309-334.

Parzen, E. (1975). Some solutions to the time series modelling and prediction problem. In *The Search for Oil* (Edited by D. Owen), 1-16. Marcel Dekker, New York.

Pötscher, B. M. and Srinivasan, S. (1994). A comparison of order estimation procedures for ARMA models. *Statist. Sinica* **4**, 29-50.

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

Rissanen, J. (1986a). Stochastic complexity and modeling. *Ann. Statist.* **14**, 1080-1100.

Rissanen, J. (1986b). Order estimation by accumulated prediction errors. In *Essays in Time Series and Applied Processes*, Special volume **23A** of *J. Appl. Probab.*, 55-61.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461-464.

Wei, C. Z. (1992). On predictive least squares principles. *Ann. Statist.* **20**, 1-42.

Department of Statistics, Stanford University, Sequoia Hall, Stanford, CA 94305, U.S.A.

(Received February 1995; accepted June 1996)