Available online at www.sciencedirect.com
ScienceDirect
Journal of Statistical Planning and Inference 137 (2007) 2802–2814
journal of statistical planning and inference
www.elsevier.com/locate/jspi

Improved model selection criteria for SETAR time series models

Pedro Galeano[a,*], Daniel Peña[b]

[a] *Departamento de Estadística e Investigación Operativa, Universidad de Santiago de Compostela, 15782 Santiago de Compostela, Spain*
[b] *Departamento de Estadística, Universidad Carlos III de Madrid, 28903 Getafe, Madrid, Spain*

Received 20 February 2006; received in revised form 18 September 2006; accepted 6 October 2006
Available online 11 February 2007

Abstract

The purpose of this paper is threefold. First, we obtain the asymptotic properties of the modified model selection criteria proposed by Hurvich et al. (1990. Improved estimators of Kullback-Leibler information for autoregressive model selection in small samples. Biometrika 77, 709–719) for autoregressive models. Second, we provide some highlights on the better performance of this modified criteria. Third, we extend the modification introduced by these authors to model selection criteria commonly used in the class of self-exciting threshold autoregressive (SETAR) time series models. We show the improvements of the modified criteria in their finite sample performance. In particular, for small and medium sample size the frequency of selecting the true model improves for the consistent criteria and the root mean square error (RMSE) of prediction improves for the efficient criteria. These results are illustrated via simulation with SETAR models in which we assume that the threshold and the parameters are unknown.

Keywords: Asymptotic efficiency; Autoregressive models; Consistency; Model selection criteria; SETAR models

1. Introduction

Since the seminal work by Akaike (1969), model selection criteria have become a widely used tool for selecting the order of different time series models. Most of these criteria can be classified into two groups. The first one includes the efficient criteria, which asymptotically select the model which produces the least mean square prediction error. The final prediction error criterion (FPE), by Akaike (1969), the Akaike information criterion (AIC), by Akaike (1973) and the corrected Akaike information criterion (AICc), by Hurvich and Tsai (1989) are efficient criteria. The FPE selects the model that minimizes the one step ahead square prediction error. The AIC is an estimator of the expected Kullback–Leibler divergence between the true and the fitted model, while the AICc is a bias correction form of the AIC that appears to work better in small samples. The second group includes the consistent criteria, which, under the assumption that the data come from a finite order autoregressive process, asymptotically select the true order of the process. The Bayesian information criterion (BIC), by Schwarz (1978), and the Hannan–Quinn criterion (HQC), by Hannan and Quinn (1979), are consistent criteria. The BIC approaches the posterior probabilities of the models, while the HQC is designed to be a consistent criterion with the fastest convergence rate to the true model.

* Corresponding author. Tel.: +34 981 563 100x13207; fax: +34 981 597 054.
E-mail address: pgaleano@usc.es (P. Galeano).

0378-3758/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2006.10.014

All these criteria can be written compactly as members of the family of criteria:

$$\min_p\{T\log\widehat{\sigma}_p^2+(p+1)\times C(T,p+1)\},\tag{1}$$

where p is the order of the autoregressive process, $\widehat{\sigma}_p^2$ is the maximum likelihood estimate of the residual variance of the process, T is the sample size and $C(T,p+1)=\frac{T}{p+1}\log(\frac{T+p+1}{T-(p+1)})$ for the FPE, $C(T,p+1)=2$ for the AIC, $C(T,p+1)=\frac{2T}{T-(p+1)-1}$ for the AICc, $C(T,p+1)=\log(T)$, for the BIC and $C(T,p+1)=2m\ \log\log(T)$ with $m>1$, for the HQC.

Hurvich et al. (1990) further approximated the expected Kullback–Leibler divergence to derive a criterion, AICc*, which can be written as follows:

$$\min_p\left\{\log|\Sigma(\widehat{\alpha}_p)|+\frac{2T(p+1)}{T-(p+1)-1}\right\},\tag{2}$$

where $|\Sigma(\widehat{\alpha}_p)|$ is the determinant of the estimated covariance matrix of the series under the autoregressive process with order p and parameters α_p, that will be defined in Section 2. These authors also introduced the AIC* and BIC* criteria, by replacing $T\log\widehat{\sigma}_p^2$ by the determinant term in the AIC and BIC criteria, and showed in a Monte Carlo experiment the good performance on this modification. However, they did not study the asymptotic properties of these modified criteria. The first contribution of this paper is to show that the asymptotic properties of the original criteria (1) applies to the modified criteria (2). Thus, we show the efficiency of AIC* and AICc* and the consistency of BIC*.

Although Hurvich et al. (1990) showed via simulation the better performance of the modified criteria, no theoretical reasons have been given explaining this improvements. The second contribution of this paper is to provide three interpretations on the advantages of using the determinant term by using three different comparisons: (1) the one step ahead prediction variances; (2) the correlation structure; (3) a measure of the goodness of the fit.

A useful nonlinear extension of linear time series models are the self-exciting threshold autoregressive (SETAR) models, see Tong (1990). These models can explain interesting features found in real data, such as asymmetric limit cycles, jump phenomena, chaos and so on. Model selection for SETAR models has been addressed in several papers. Tong (1990) suggested to use the AIC but no theoretical justification was given. Wong and Li (1998) showed that the AICc criterion is an asymptotically unbiased estimator of the expected Kullback–Leibler information for SETAR models and analyzed the small sample properties of AIC, AICc and BIC via simulation experiments. Kapetanios (2001) extended some of the existing theoretical results for several model selection criteria in linear models to threshold models. De Gooijer (2001) proposed three cross-validation criteria. Campbell (2004) and Unnikrishnan (2004) developed Bayesian model selection within a Markov Chain Monte Carlo (MCMC) framework, and, finally, Öhrvik and Schoier (2005) studied the performance of several bootstrap selection criteria.

As a SETAR model is piecewise autoregressive linear, it seems natural to extend the modification considered by Hurvich et al. (1990) for autoregressive models to these nonlinear models. The third contribution of this paper is to present new SETAR model selection criteria based on the determinant term and show via a Monte Carlo study the better performance for small and medium sample size of these modified criteria.

The rest of this paper is organized as follows. Section 2 briefly reviews model selection criteria for the class of linear autoregressive models, proves that the correction by the determinant term keeps their asymptotic properties, and provides some intuition to justify why this correction can improve their performance. Section 3 develops the modification by the determinant term for SETAR time series model selection criteria. Section 4 shows the better performance of the modified criteria in a Monte Carlo experiment.

2. Model selection criteria for the class of linear autoregressive processes

2.1. Model selection criteria for autoregressive processes

Suppose it is known that a given time series, $x=(x_1,\ldots,x_T)'$, has been generated by the class of autoregressive (AR) Gaussian processes, given by

$$x_t-\phi_1x_{t-1}-\cdots-\phi_px_{t-p}=a_t,\quad t=\ldots,-1,0,1,\ldots,$$

where a_t is a sequence of independent Gaussian distributed random variables with zero mean and variance σ_p^2. We assume that $p \in \{0, \ldots, p_{\max}\}$, where $p_{\max}$ is an upper bound. The AR(p) model, denoted by M_p, has parameters $\alpha_p = (\beta_p', \sigma_p^2)'$, where β_p is the $p_{\max} \times 1$ vector of parameters,

$$\beta_p = \left(\underbrace{\phi_{1p}, \ldots, \phi_{pp}, 0, \ldots, 0}_{1 \times p_{\max}} \right)'.$$

We assume that the models M_p are causal, invertible and stationary. Then the covariance matrix of x can be written as $\Sigma(\alpha_p) = \sigma_p^2 Q(\beta_p)$, where $Q(\beta_p)$ is a $T \times T$ matrix which only depends on β_p. Let,

$$Q(\beta_p) = L(\beta_p) L'(\beta_p)$$

be the Cholesky decomposition of $Q(\beta_p)$ such that $a(\beta_p) = L(\beta_p)^{-1} x$. We denote the parameters of the model that have generated the data as $\alpha_0 = (\beta_0', \sigma_0^2)'$.

In practice, the model parameters are unknown. The maximum likelihood estimates of the vector of parameters β_p and the innovations variance σ_p^2, denoted by $\widehat{\beta}_p$ and $\widehat{\sigma}_p^2$, respectively, are obtained after maximizing the likelihood function, $p(x|M_p)$, given by:

$$p(x|M_p) = (2\pi)^{-T/2} |\Sigma(\alpha_p)|^{-1/2} \exp\left(-\frac{1}{2} x' \Sigma^{-1}(\alpha_p) x \right). \tag{3}$$

Akaike (1973) proposed to select the model which minimizes the expected Kullback–Leibler divergence between the fitted and the true model, defined by,

$$E_{\widehat{\alpha}_p}\left[E_{\alpha_0}\left[2 \log \frac{p(y|\alpha_0)}{p(y|\widehat{\alpha}_p)} \right] \right] = E_{\widehat{\alpha}_p}[E_{\alpha_0}[-2 \log p(y|\widehat{\alpha}_p)]] - E_{\widehat{\alpha}_p}[E_{\alpha_0}[-2 \log p(y|\alpha_0)]] \tag{4}$$

for an arbitrary realization $y = (y_1, \ldots, y_T)'$ of the process. As (4) is always positive, minimizing it implies making $p(y|\widehat{\alpha}_p)$ as close as possible to $p(y|\alpha_0)$, in the expected Kullback–Leibler divergence. As the second term of the expected Kullback–Leibler divergence is constant for all the models, minimizing (4) is equivalent to minimize,

$$E_{\widehat{\alpha}_p}[E_{\alpha_0}[-2 \log p(y|\widehat{\alpha}_p)]] = \int \left[\int -2 \log p(y|\widehat{\alpha}_p) p(y|\alpha_0) \, \mathrm{d}y \right] p(\widehat{\alpha}_p|\alpha_0) \, \mathrm{d}\widehat{\alpha}_p, \tag{5}$$

where y and $\widehat{\alpha}_p$ are assumed to be independent. Thus, the rule proposed by Akaike selects the autoregressive model that minimizes (5) with respect to the two sources of uncertainty: the distribution of future observations given the parameters and the distribution of the estimate. Akaike (1973) approached (5) as follows:

$$E_{\widehat{\alpha}_p}[E_{\alpha_0}[-2 \log p(y|\widehat{\alpha}_p)]] = T(\log 2\pi + 1) + T \log \widehat{\sigma}_p^2 + 2(p+1) + \mathrm{o}_p(1),$$

which leads to the AIC

$$\mathrm{AIC}(p) = T \log \widehat{\sigma}_p^2 + 2(p+1).$$

Hurvich and Tsai (1989) obtained an approximation of (5) which reduces the small sample bias of the approximation by Akaike (1973), and is given by

$$E_{\widehat{\alpha}_p}[E_{\alpha_0}[-2 \log p(y|\widehat{\alpha}_p)]] = T(\log 2\pi + 1) + T \log \widehat{\sigma}_p^2 + \frac{2T(p+1)}{T-(p+1)-1} + \mathrm{o}_p(1)$$

which leads to the AICc

$$\mathrm{AICc}(p) = T \log \widehat{\sigma}_p^2 + \frac{2T(p+1)}{T-(p+1)-1}.$$

From the Bayesian point of view, the model selected is the one with maximum posterior probability, $p(M_p|x)$, where $p(M_p|x) \propto p(x|M_p) p(M_p)$, and $p(x|M_p) = \int p(x|\alpha_p, M_p) p(\alpha_p|M_p) \, \mathrm{d}\alpha_p$, such that $p(M_p)$ and $p(\alpha_p|M_p)$ are the

prior probabilities of models and parameters, respectively. Schwarz (1978) approximated the posterior probability $p(M_p|x)$ to derive the BIC, given by

$$\mathrm{BIC}(p) = T\log\widehat{\sigma}_p^2 + \log(T)(p+1).$$

The criteria AIC, AICc and BIC can be written in a compact way as members of the family of criteria,

$$\min_p\{T\log\widehat{\sigma}_p^2 + (p+1)\times C(T, p+1)\}, \tag{6}$$

where $C(T, p+1)$ is 2, for AIC, $\frac{2T}{T-(p+1)-1}$, for AICc, and $\log(T)$, for BIC.

Hurvich et al. (1990) noted that the expected Kullback–Leibler divergence was better approximated if $T\log\widehat{\sigma}_p^2$ is replaced by the term $\log|\Sigma(\widehat{\alpha}_p)|$ and defined the AIC$^*(p)$, AICc$^*(p)$ and BIC$^*(p)$ criteria, which may be written in compact form as:

$$\min_p\{\log|\Sigma(\widehat{\alpha}_p)| + (p+1)\times C(T, p+1)\}. \tag{7}$$

These authors showed by simulation the advantages of considering $\log|\Sigma(\widehat{\alpha}_p)|$ instead of $T\log\widehat{\sigma}_p^2$ for autoregression fitting, but did not analyze the asymptotic properties of the modified criteria. First, we show that the AIC* and AICc* criteria are efficient in the following Theorem, whose proof is in Appendix A.

Theorem 1. *Assume that the following assumptions hold*: (A1) $\{x_t\}$ is *generated by a stationary process* $x_t-\phi_1x_{t-1}-\phi_2x_{t-2}-\cdots=a_t, t=\cdots,-1,0,1,\cdots$ *where* a_t *is a sequence of independent Gaussian distributed random variables with zero mean and variance* σ_a^2 *and* $\sum_{j=1}^{\infty}|\phi_j|<\infty$; (A2) *the polynomial* $\phi(z)=1-\phi_1z-\phi_2z^2-\cdots$, *is nonzero for every complex number* z *with* $|z|\leqslant 1$; (A3) *the upper bound* $p_{\max}$ *is a sequence of positive integers which depends on* T *such that* $p_{\max}\to\infty$ *and* $p_{\max}/\sqrt{T}\to 0$ *as* $T\to\infty$; (A4) $\{x_t\}$*is not degenerate to a finite order autoregressive process. Then, the* AIC* *and* AICc* *are efficient criteria.*

On the other hand, the consistency property of the BIC* criterion is established in the following Theorem, proved in Appendix A.

Theorem 2. *Assume that the following assumptions hold*: (B1) $\{x_t\}$ *is generated by a stationary process* $x_t-\phi_1x_{t-1}-\cdots-\phi_px_{t-p}=a_t, t=\cdots,-1,0,1,\ldots$ *where* a_t *is a sequence of independent Gaussian distributed random variables with zero mean and variance* σ_a^2; (B2) *the polynomial* $\phi(z)=1-\phi_1z-\phi_2z^2-\cdots-\phi_pz^p$, *is nonzero for every complex number* z *with* $|z|\leqslant 1$; (B3) *the upper bound* $p_{\max}$ *is fixed and known a priori. Then, the* BIC* *is a consistent criteria.*

Thus, AIC*, AICc* and BIC* are similar to AIC, AICc and BIC for large samples but as we will see in the Monte Carlo study in Section 4, in small and medium sample settings, the difference between the performance of these criteria may be substantial.

2.2. *Three interpretations on the advantages of using* AIC*, AICc* *and* BIC*

Next, we provide three interpretations on the advantages of using the determinant term $|\Sigma(\widehat{\alpha}_p)|$ by showing that this term leads to a better comparison of the models under consideration by (1) the one step ahead prediction variances; (2) the correlation structure; (3) a goodness of fit test.

2.2.1. *Interpretation by one step ahead prediction variances*

To show the first interpretation, let $\widehat{x}_t(p)$ be the one step ahead mean square predictions under the M_p model and let $e_t(p)=x_t-\widehat{x}_t(p)$ be the corresponding one step ahead prediction errors. These errors have variances which can be written as $E[e_t(p)^2]=\sigma_p^2v_t^2(p)$ (see, Harvey, 1981). For instance, for the AR(1) model, $v_t^2(1)=1/(1-\phi^2)$, for $t=1$,

and $v_t^2(p) = 1$, for $t > 1$. Thus, the logarithm of the determinant of the covariance matrix of x can be written as

$$\log|\Sigma(\widehat{\alpha}_p)| = T\log\left(\prod_{t=1}^{T}\widehat{\sigma}_p^2\widehat{v}_t^2(p)\right)^{1/T} = \sum_{t=1}^{T}\log\widehat{\sigma}_p^2\widehat{v}_t^2(p),$$

where $\widehat{v}_t^2(p)$ are obtained after replacing the estimated parameters in $v_t^2(p)$, and $\log|\Sigma(\widehat{\alpha}_p)|$ is T times the logarithm of the geometric mean of the estimated one step ahead prediction variances. Therefore, the difference $\mathrm{AIC}^*(p+1) - \mathrm{AIC}^*(p)$ can be written as

$$\mathrm{AIC}^*(p+1) - \mathrm{AIC}^*(p) = \sum_{t=1}^{T}\log\frac{\widehat{\sigma}_{p+1}^2\widehat{v}_t^2(p+1)}{\widehat{\sigma}_p^2\widehat{v}_t^2(p)} + 2.$$

The first term measures the relative change between the one step ahead prediction variances, while the second term is a penalization for the inclusion of one additional parameter. Therefore, $\mathrm{AIC}^*(p+1) < \mathrm{AIC}^*(p)$ if the geometric mean of the one step ahead prediction variances under the model M_{p+1} is significatively smaller than the corresponding mean under the model M_p, or in other words, the AIC* will select the model which has a better predictive performance penalized by the number of parameters. As

$$\mathrm{AIC}^*(p) = T\log\widehat{\sigma}_p^2 + \sum_{t=1}^{T}\log\widehat{v}_t^2(p) + 2(p+1) = \mathrm{AIC}(p) + \sum_{t=1}^{T}\log\widehat{v}_t^2(p),$$

the $\mathrm{AIC}(p)$ does not take into account the terms $\widehat{v}_t^2(p)$, but only the estimated residual variance, $\widehat{\sigma}_p^2$. The same conclusions holds for the AICc* and BIC* criteria.

2.2.2. Interpretation by the correlation structure

Now we show that the determinant term provides a more sophisticated comparison of the autocorrelation structure of the models under consideration. As $Q(\beta_p) = (\sigma_x^2/\sigma_p^2)R(\alpha_p)$, where σ_x^2 is the variance of x and $R(\alpha_p)$ is the correlation matrix of x, we have $|Q(\beta_p)| = (\sigma_x^2/\sigma_p^2)^{\mathrm{T}}|R(\alpha_p)|$. Durbin (1960) and Ramsey (1974) showed that

$$\sigma_x^2/\sigma_p^2 = \prod_{i=1}^{p}(1-\phi_{ii}^2(\beta_p))^{-1}, \quad |R(\alpha_p)| = \prod_{i=1}^{p}(1-\phi_{ii}^2(\beta_p))^{T-i},$$

respectively, where $\phi_{ii}(\beta_p)$ are the partial autocorrelations of the process under the model M_p, so that,

$$|Q(\beta_p)| = \frac{\prod_{i=1}^{p}(1-\phi_{ii}^2(\beta_p))^{T-i}}{\prod_{i=1}^{p}(1-\phi_{ii}^2(\beta_p))^{T}} = \prod_{i=1}^{p}(1-\phi_{ii}^2(\beta_p))^{-i}.$$

Consequently, the criteria (7) can be written as follows:

$$\min_p\left\{T\log\widehat{\sigma}_p^2 + (p+1)\times C(T,p+1) - \sum_{i=1}^{p} i\log(1-\phi_{ii}^2(\widehat{\beta}_p))\right\},$$

while the difference $\mathrm{AIC}^*(p+1) - \mathrm{AIC}^*(p)$ is

$$\mathrm{AIC}^*(p+1) - \mathrm{AIC}^*(p) = T\log\frac{\widehat{\sigma}_{p+1}^2}{\widehat{\sigma}_p^2} + 2 - \sum_{i=1}^{p} i\log\frac{(1-\phi_{ii}^2(\widehat{\beta}_{p+1}))}{(1-\phi_{ii}^2(\widehat{\beta}_p))} - (p+1)\log(1-\phi_{p+1,p+1}^2(\widehat{\beta}_{p+1})).$$

The first two terms are the one used by the AIC criterion but now two additional terms appear in the comparison. They measure the discrepancy between all the partial autocorrelation coefficients under both hypothesis, M_p and M_{p+1}, with weights that increase with the lag. Therefore, $\mathrm{AIC}^*(p+1) < \mathrm{AIC}^*(p)$ if either: (a) $\widehat{\sigma}_{p+1}^2$ is smaller enough than $\widehat{\sigma}_p^2$

or (b) the weighted sum of the partial autocorrelation coefficients computed under the AR(p) model is greater enough than the corresponding sum under the M_{p+1} model. Note that the last term is acting as a penalization term because it is always positive. The same interpretation applies to BIC* and AICc*, and the only difference is the penalization term for including one additional parameter.

2.2.3. Interpretation by a goodness of fit test

A last interpretation is given in terms of goodness of fit tests for time series. Peña and Rodriguez (2006) used the log of the determinant of the autocorrelation matrix of the estimated residuals for testing goodness of fit in time series. As $a(\beta_p) = L(\beta_p)^{-1}x$, we can write

$$\frac{1}{T}a(\beta_p)a(\beta_p)' = [L(\beta_p)^{-1}]\left[\frac{1}{T}xx'\right][L(\beta_p)^{-1}]'. \tag{8}$$

Thus, after fitting the model M_p,

$$\frac{1}{T}a(\widehat{\beta}_p)a(\widehat{\beta}_p)' = \widehat{\sigma}_p^2 R(a(\widehat{\beta}_p)), \tag{9}$$

where $R(a(\widehat{\beta}_p))$ is the sample correlation matrix of the estimated residuals, $a(\widehat{\beta}_p)$. Therefore, from (8) and (9), we have

$$T\log\widehat{\sigma}_p^2 + \log|Q(\widehat{\beta}_p)| = \log|(1/T)xx'| - \log|R(a(\widehat{\beta}_p))|$$

so that,

$$\mathrm{AIC}^*(p+1) - \mathrm{AIC}^*(p) = \log|R(a(\widehat{\beta}_p))| - \log|R(a(\widehat{\beta}_{p+1}))| + 2.$$

Thus, $\mathrm{AIC}^*(p+1) < \mathrm{AIC}^*(p)$ if the logarithm of the determinant of the correlation matrix of the estimated residuals after the model M_{p+1} is significative larger than the one for the model M_p. The terms $\log|R(a(\widehat{\beta}_p))|$ and $\log|R(a(\widehat{\beta}_{p+1}))|$ are the values of the statistic proposed by Peña and Rodriguez (2006) for models M_p and M_{p+1}. Therefore, the AIC* will select the model with have a significatively larger value of the statistic proposed by Peña and Rodriguez (2006). Consequently, the term $T\log\widehat{\sigma}_p^2 + \log|Q(\widehat{\beta}_p)|$, can be seen as a measure of the goodness of fit of the model M_p to the series x. As before, the same interpretation applies to BIC* and AICc* after changing the penalization term.

3. Model selection in SETAR models

One of the most often used nonlinear time series model is the SETAR model. A time series data, $x = (x_1, \ldots, x_T)'$, generated by the class of SETAR processes follows the model:

$$x_t = \phi_{j0} + \sum_{i=1}^{p_j}\phi_{ji}x_{t-i} + a_{jt}, \quad \text{if } r_{j-1} \leqslant x_{t-d} < r_j, \quad j = 1, \ldots, k, \tag{10}$$

where a_{jt} are sequences of independent Gaussian distribution random variables with zero mean and variances σ_j^2. We assume that $p_j \in \{0, \ldots, p_j^{\max}\}$ and $d \in \{0, \ldots, d_{\max}\}$ are nonnegative integers, where $p_j^{\max}$, $j = 1, \ldots, k$ and $d_{\max}$, are some upper bounds, and $-\infty = r_0 < r_1 < \cdots < r_{k-1} < r_k = \infty$ are the thresholds. The SETAR($p_1, \ldots, p_k, d$) model, denoted by $M_{p_1,\ldots,p_k,d}$, has parameters $\alpha_{p_1,\ldots,p_k,d} = (\beta'_{p_1,\ldots,p_k,d}, \sigma_1^2, \ldots, \sigma_{p_k}^2)'$, where $\beta_{p_1,\ldots,p_k,d}$ is the $(\sum_{j=1}^k (p_j^{\max} + 1) + k + 2) \times 1$ vector of parameters,

$$\beta_{p_1,\ldots,p_k,d} = \left(\underbrace{\phi_{10}, \phi_{11}, \ldots, \phi_{1p_1}, 0, \ldots, 0, \phi_{k0}, \phi_{k1}, \ldots, \phi_{kp_k}, 0, \ldots, 0}_{1\times\sum_{j=1}^k (p_j^{\max}+1)}, \underbrace{r_0, \ldots, r_k}_{1\times(k+1)}, d\right)'.$$

We assume that the models $M_{p_1,\ldots,p_k,d}$ are stationary, ergodic with finite second moments and the stationary distribution of $x=(x_1,\ldots,x_T)'$ admits a density that is positive everywhere. We denote the parameters of the model that have generated the data as α_0.

Exact maximum likelihood estimates of the parameters of model (10) are not considered because of the complexity of the likelihood function. Assuming that d is known, the conditional log-likelihood of model (10) is given by

$$\log p_c(x|\alpha_{p_1,\ldots,p_k,d}) = -\frac{T}{2}\log(2\pi) - \frac{1}{2}\sum_{j=1}^{k}\left(T_j \log\sigma_{p_j}^2 + \frac{S(\phi_j, r_{j-1}, r_j)}{\sigma_{p_j}^2}\right), \tag{11}$$

where T_j are the number of observations in each regime for the thresholds $r_1,\ldots,r_{k-1}$, $\phi_j=(\phi_{j0},\ldots,\phi_{jp_j})'$ and $S(\phi_j, r_{j-1}, r_j)=\sum_{r_{j-1}\leqslant x_{t-d}<r_j} a_{jt}^2$, $j=1,\ldots,k$. The conditional maximum likelihood estimators of the parameters, denoted by $\widehat{\alpha}_{p_1,\ldots,p_k,d}=(\widehat{\beta}'_{p_1,\ldots,p_k,d}, \widehat{\sigma}_1^2,\ldots,\widehat{\sigma}_{p_k}^2)'$, are the values that maximize the conditional likelihood in (11), with residual variances

$$\widehat{\sigma}_j^2 = \frac{S(\widehat{\phi}_j, \widehat{r}_{j-1}, \widehat{r}_j)}{T_j}, \quad j=1,\ldots,k.$$

Chan (1993) showed the strong consistency of the conditional least squares estimators of the parameters for $k=1$.

Wong and Li (1998) approximated the expected Kullback–Leibler divergence for SETAR models as follows:

$$E_{\widehat{\alpha}_{p_1,\ldots,p_k,d}}[E_{\alpha_0}[-2\log p_c(y|\widehat{\alpha}_{p_1,\ldots,p_k,d})]] = T\log(2\pi) + \sum_{i=1}^{k} T_i \log\widehat{\sigma}_{p_i}^2 + \sum_{i=1}^{k}\frac{T_i(T_i+p_i+1)}{(T_i-p_i-3)} + o_p(1),$$

which leads to the AICc criterion for SETAR models. These authors compared in a simulation study three model selection criteria, AIC, AICc and BIC, which for k regimes are given by

$$\min_{(p_1,\ldots,p_k)}\left\{\sum_{j=1}^{k}[T_j\log\widehat{\sigma}_j^2 + (p_j+1)\times C_j(T_j, p_j+1)]\right\}, \tag{12}$$

where $C_j(T_j, p_j+1)$ is 2 for AIC, $\frac{1}{p_j+1}\frac{T_j(T_j+p_j+1)}{T_j-(p_j+1)-2}$ for AICc, and $\log T_j$ for BIC. The procedure proposed by Wong and Li (1998) works as follows, when $k=2$, $r_1=r$ and d are unknown: (a) fix the maximum autoregressive and delay orders $\{p_1^{\max}, p_2^{\max}, d^{\max}\}$; (b) assume $r\in[l,u]\subset R$, where l is the $0.25\times 100\%$ percentile and u is the $0.75\times 100\%$ percentile of x; (c) let $x_{(1)},\ldots,x_{(T)}$ be the order statistics of x; (d) let $I_r=\{[0.25T],\ldots,[0.75T]\}$, where $[vT]$ is the largest integer less than vT. Set $r=x_{(i)}$, $i\in I_r$; (e) calculate

$$\min_{(p_1,p_2,d,x_{(i)})}\{\mathrm{MSC}(p_1, p_2, d, x_{(i)})\},$$

where $\mathrm{MSC}(p_1, p_2, d, x_{(i)})$ is one of the model selection criteria in (12). The autoregressive orders (p_1, p_2), the delay parameter, d, and the threshold, $x_{(i)}$, selected are the ones that minimize $\mathrm{MSC}(p_1, p_2, d, x_{(i)})$. Wong and Li (1998) carried out a Monte Carlo experiment for different models and sample sizes for the criteria in (12), and conclude that the AICc has the best performance for small sample sizes and BIC for medium and large sample sizes.

De Gooijer (2001) proposed a procedure for selecting and estimating the parameters of a SETAR model by cross-validation which works as follows: (a)–(d) are as in the previous procedure; (e) omit one observation of the series, x_t, and with the remaining data set obtain conditional least squares estimates of the parameters of the corresponding model, which we denote by $\widehat{\phi}_j^t$, predict the omitted observation and obtain the predictive residual, $a_t(\widehat{\phi}_j^t, \widehat{r}_{j-1}, \widehat{r}_j)$; (f) repeat the previous step for all the observations. The final model is the one that minimizes one of the criteria C_1, Cc and

Cu, written in compact way as follows:

$$\min_{(p_1,\ldots,p_k)} \left\{ T \log\left(\frac{1}{T} \sum_{t=1}^{T} a_t^2(\widehat{\phi}_j^t, \widehat{r}_{j-1}, \widehat{r}_j) \right) + \sum_{j=1}^{k} (p_j+1) \times C_j(T_j, p_j+1) \right\}, \tag{13}$$

where $C_j(T_j, p_j+1)$ is 0 for C_1, $\frac{1}{p_j+1}\frac{T_j(T_j+p_j+1)}{T_j-p_j-3}$ for Cc, and $\frac{1}{p_j+1}[\frac{T_j(T_j+p_j+1)}{T_j-p_j-3} + T_j \log\{\frac{T_j}{T_j-p_j-2}\}]$ for Cu. The C_1 criterion was analyzed in Stoica et al. (1986) for linear models and proved that for a given model, $C_1 = \text{AIC} + O(T^{-1/2})$. The Cc and the Cu criteria came from adding the penalty terms of the AICc and AICu criteria to the C_1 criterion. The AICu is a criterion introduced by McQuarrie et al. (1997) for linear models and is neither efficient nor consistent but it has a good performance in finite samples. For SETAR models the AICu criterion can be written as in (12) with $C_j(T_j, p_j+1) = \frac{1}{p_j+1}[\frac{T_j(T_j+p_j+1)}{T_j-p_j-3} + T_j \log\{\frac{T_j}{T_j-p_j-2}\}]$.

As SETAR models are piecewise autoregressive linear, we propose to modify the criteria as in the autoregressive models case as follows. The criteria BIC, AIC, AICc and AICu are modified by adding the determinant term in each regime as follows:

$$\min_{(p_1,\ldots,p_k)} \left\{ \sum_{j=1}^{k} [T_j \log \widehat{\sigma}_j^2 + (p_j+1) \times C_j(T_j, p_j+1) + \log|Q(\widehat{\phi}_j)|] \right\}. \tag{14}$$

In order to compute the determinant term in each regime we first estimate the parameters of the model by conditional likelihood and then obtain the determinant term in each regime. For that we use the expression provided by Leeuw (1994) who showed that

$$|Q(\widehat{\phi}_j)| = \frac{1}{|M'M - NN'|}, \tag{15}$$

where M and N are $p_j \times p_j$ matrices with elements given by

$$M_{ab} = \begin{cases} 0 & a < b, \\ 1 & a = b, \\ -\widehat{\phi}_{j,a-b} & a > b, \end{cases} \qquad N_{ab} = \begin{cases} -\widehat{\phi}_{j,p_j+(a-b)}, & a \leqslant b, \\ 0, & a > b. \end{cases}$$

In the same way, we modify the cross-validation criteria C_1, Cc and Cu proposed by De Gooijer (2001) in (13) by adding the determinant term in each regime. Therefore, the modified cross-validation criteria C_1^*, Cc* and Cu* are defined as follows:

$$\min_{(p_1,\ldots,p_k)} \left\{ T \log\left(\frac{1}{T} \sum_{t=1}^{T} a_t^2(\widehat{\phi}_j^t, \widehat{r}_{j-1}, \widehat{r}_j) \right) + \sum_{j=1}^{k} [(p_j+1) \times C_j(T_j, p_j+1) + \log|Q(\widehat{\phi}_j)|] \right\}. \tag{16}$$

The procedures in Wong and Li (1998) and De Gooijer (2001) are modified by adding the determinant term in the last step obtained with the conditional least squares estimates of the parameters with the whole series. Then, the final model selected is the one that minimizes one of the criteria in (16). The determinant term is computed using (15).

4. Monte Carlo experiments

To evaluate the performance of the proposed criteria for SETAR models, 1000 realizations were generated from the following two stationary SETAR models:

$$\text{(M1)} \begin{cases} x_t = -0.8x_{t-1} + a_{1t}, & x_{t-1} \leqslant 0, \\ x_t = -0.2x_{t-1} + a_{2t}, & x_{t-1} > 0, \end{cases} \qquad \text{(M2)} \begin{cases} x_t = 0.5x_{t-1} + a_{1t}, & x_{t-1} \leqslant 0, \\ x_t = -0.5x_{t-1} + a_{2t}, & x_{t-1} > 0, \end{cases}$$

Table 1
Frequency of times of correct selection, root mean square errors of the threshold parameter and root mean square prediction errors assuming that d is known

M	$T=30$	BIC	BIC*	AIC	AIC*	AICc	AICc*	AICu	AICu*	C_1	C_1^*	Cc	Cc*	Cu	Cu*
1	(p_1, p_2)	306	377	254	331	800	856	903	923	363	474	818	854	895	923
1	*RMSE*	0.62	0.62	0.60	0.60	0.41	0.40	0.40	0.39	0.55	0.54	0.40	0.40	0.40	0.39
1	*RMSPE*	1.41	1.37	1.38	1.32	1.13	1.13	1.12	1.12	1.22	1.22	1.14	1.12	1.12	1.11
2	(p_1, p_2)	306	361	243	310	779	825	876	913	378	450	786	840	890	921
2	*RMSE*	0.84	0.84	0.82	0.82	0.65	0.66	0.65	0.65	0.73	0.74	0.64	0.65	0.63	0.63
2	*RMSPE*	1.64	1.63	1.64	1.64	1.16	1.15	1.13	1.12	1.31	1.31	1.14	1.13	1.12	1.12
	$T=50$														
1	(p_1, p_2)	420	512	198	298	629	686	789	827	286	395	634	688	796	836
1	*RMSE*	0.58	0.56	0.55	0.54	0.47	0.47	0.47	0.47	0.53	0.52	0.45	0.45	0.45	0.44
1	*RMSPE*	1.15	1.14	1.19	1.19	1.06	1.06	1.06	1.06	1.15	1.15	1.07	1.07	1.07	1.06
2	(p_1, p_2)	412	507	216	307	633	704	802	846	313	406	659	724	820	855
2	*RMSE*	0.74	0.74	0.73	0.74	0.63	0.63	0.63	0.64	0.70	0.70	0.63	0.65	0.64	0.64
2	*RMSPE*	1.19	1.19	1.24	1.23	1.15	1.15	1.13	1.12	1.22	1.22	1.15	1.14	1.12	1.12
	$T=100$														
1	(p_1, p_2)	743	796	359	431	537	591	753	789	358	439	522	589	765	799
1	*RMSE*	0.54	0.52	0.52	0.51	0.50	0.48	0.50	0.47	0.51	0.49	0.49	0.47	0.48	0.46
1	*MSPE*	1.06	1.06	1.08	1.06	1.07	1.06	1.06	1.06	1.08	1.08	1.07	1.07	1.06	1.06
2	(p_1, p_2)	771	819	365	446	545	600	773	816	398	461	542	611	773	805
2	*RMSE*	0.59	0.59	0.58	0.59	0.57	0.58	0.57	0.57	0.58	0.58	0.56	0.57	0.56	0.57
2	*RMSPE*	1.06	1.05	1.07	1.07	1.07	1.07	1.06	1.05	1.07	1.07	1.07	1.06	1.06	1.06

where $a_{jt} \sim \mathrm{N}(0, 1)$, $j = 1, 2$. Based on Section 3, we compare the performance of the criteria in (12) with respect to the criteria in (14) and the criteria in (13) with respect to the criteria in (16). In all cases, 1000 series were generated from models M1 and M2 with sample sizes $T = 31$, 51 and 101. We proceed as in Wong and Li (1998) and De Gooijer (2001) by using a grid to estimate the threshold parameter r. We fit each model to the first $T - 1$ observations of each series by conditional likelihood and obtain the determinant term in (15) in each regime. First, we assume that the delay parameter is known and fix $p_1^{\max} = p_2^{\max} = 5$ for $T = 31$, 51 and 101, so that taking into account that the number of possible values of the threshold parameter is $(T - 1)/2$, we compare 375, 625 and 1250 models, respectively. In every case, we consider the following measures of the performance of the model selection criteria: (a) the frequency detection of the correct order $(p_1, p_2) = (1, 1)$; (b) the root mean square error (RMSE) of estimation of the threshold parameter and (c) the root mean square prediction error (RMSPE) for the last observation by using the model chosen by each criteria, the fitted parameters and the true value. The results are in Table 1. It can be seen that for small sample size, $T = 30$, the improvement in the number of times in which the correct model is selected can be as large as 30.5% (see C_1 and C_1^* in M1), for $T = 50$ as large as 50.5% (see AIC and AIC* in M1) and for $T = 100$ as large as 22.6% (see AIC and AIC* in M1). First part of Table 3 includes the improvement percentage of the number of times in which the correct orders are selected by all the modified criteria. We note that the AICu, AICu*, Cu and Cu* have larger frequency detection for $T = 30$ but the frequency detection decreases when the sample size increases. On the other hand, the RMSE of estimation of the threshold parameter are very close for the original and modified criteria, whereas the RMSPE is usually smaller for the modified criteria.

Now, we assume that the delay is unknown and apply the same design as before with $p_1^{\max} = p_2^{\max} = 5$, $d^{\max} = 4$ and $T = 31$, 51 and 101 . Now we compare 1500, 2500 and 5000 models, respectively. In every case, we consider the same three measures of the performance of the model selection criteria, the frequency detection of the correct order (p_1, p_2), RMSE and RMSPE but we also include the frequency detection of selecting the correct delay parameter (d) and the frequency detection of the correct orders and delay parameter, (p_1, p_2, d). The results are given in Table 2.

Table 2
Frequency of times of correct selection, root mean square errors of the threshold parameter and root mean square prediction errors assuming that d is unknown

M	$T=30$	BIC	BIC*	AIC	AIC*	AICc	AICc*	AICu	AICu*	C_1	C_1^*	Cc	Cc*	Cu	Cu*
1	(p_1, p_2)	205	303	163	248	739	808	862	897	292	403	765	824	863	913
1	(d)	540	559	553	571	582	616	589	617	540	540	573	578	573	575
1	(p_1, p_2, d)	135	187	117	165	463	517	527	566	187	240	469	499	517	536
1	*RMSE*	0.65	0.64	0.63	0.62	0.43	0.42	0.42	0.41	0.55	0.55	0.41	0.41	0.41	0.41
1	*RMSPE*	1.62	1.61	1.64	1.62	1.18	1.17	1.17	1.17	1.29	1.28	1.19	1.19	1.18	1.18
2	(p_1, p_2)	184	249	147	221	733	802	846	879	292	390	754	809	860	885
2	(d)	569	555	562	568	601	590	608	598	611	599	649	632	653	633
2	(p_1, p_2, d)	119	146	98	141	461	485	527	538	203	254	512	526	577	570
2	*RMSE*	0.90	0.90	0.88	0.88	0.68	0.69	0.67	0.67	0.78	0.78	0.67	0.68	0.66	0.67
2	*RMSPE*	1.90	1.86	1.90	1.86	1.27	1.25	1.24	1.22	1.52	1.51	1.28	1.27	1.25	1.24
	$T=50$														
1	(p_1, p_2)	223	340	85	168	529	624	744	800	187	292	554	645	746	812
1	(d)	335	337	337	320	364	365	369	371	384	380	394	403	406	419
1	(p_1, p_2, d)	99	130	40	63	225	250	296	318	107	149	252	288	326	360
1	*RMSE*	0.63	0.61	0.61	0.60	0.48	0.47	0.47	0.47	0.53	0.53	0.46	0.46	0.45	0.45
1	*RMSPE*	1.25	1.22	1.30	1.26	1.17	1.16	1.14	1.14	1.24	1.23	1.18	1.18	1.16	1.15
2	(p_1, p_2)	247	325	105	183	570	655	773	824	217	293	594	668	784	834
2	(d)	419	421	400	423	483	474	500	483	480	478	528	521	544	529
2	(p_1, p_2, d)	146	182	66	115	325	347	419	427	143	178	359	380	455	460
2	*RMSE*	0.84	0.84	0.82	0.82	0.68	0.68	0.68	0.68	0.72	0.72	0.66	0.66	0.65	0.66
2	*RMSPE*	1.44	1.44	1.39	1.35	1.20	1.20	1.17	1.16	1.31	1.29	1.20	1.19	1.18	1.18
	$T=100$														
1	(p_1, p_2)	652	747	221	330	388	480	667	773	235	316	401	481	662	773
1	(d)	491	542	489	527	504	542	522	527	552	567	567	572	582	582
1	(p_1, p_2, d)	351	421	135	192	210	286	376	436	195	251	301	351	436	481
1	*RMSE*	0.53	0.52	0.52	0.50	0.49	0.46	0.49	0.47	0.49	0.49	0.48	0.47	0.49	0.46
1	*MSPE*	1.04	1.03	1.05	1.05	1.06	1.06	1.04	1.02	1.06	1.06	1.04	1.02	1.04	1.03
2	(p_1, p_2)	808	888	371	466	547	632	838	863	421	532	602	662	798	863
2	(d)	732	727	667	667	677	692	717	712	747	773	773	788	788	793
2	(p_1, p_2, d)	632	662	291	371	431	486	632	637	376	456	517	562	657	697
2	*RMSE*	0.58	0.59	0.60	0.59	0.56	0.57	0.56	0.56	0.56	0.56	0.54	0.57	0.55	0.56
2	*RMSPE*	1.04	1.03	1.05	1.05	1.04	1.04	1.04	1.03	1.06	1.06	1.07	1.05	1.05	1.05

It can be seen that for small sample size, $T=30$, the improvement in the number of times in which the correct orders $(p_1, p_2, d)=(1, 1, 1)$ are selected can be as large as 43.8% (see AIC and AIC* in M2), for $T=50$ as large as 74.2% (see AIC and AIC* in M2) and for $T=100$ as large as 42.2% (see AIC and AIC* in M1). See the second part of Table 3 to find the improvement percentage of the number of times in which the correct orders and delay parameter are selected by the modified criteria. As in the case in which d is assumed known, the AICu, AICu*, Cu and Cu* have the larger frequency detection for the true autoregression orders and delay parameters for $T=30$, but the frequency detection decreases when the sample size increases. We note that sometimes the modified criteria have a shorter frequency detection of the delay parameter, but this is not a drawback for them because the aim is to detect both the true autoregressive orders and the delay parameter, and not only the delay parameter. Regarding the RMSE and the RMSPE, the results are similar to the case in which d is assumed to be known. In terms of computational effort, the time needed to select the model by all the criteria considered in this paper for a series with $T=31$, 51 and 101, when both the threshold and the delay parameter are unknown are 21.8, 81.9 and 547 s, respectively, using a program written in Matlab with a Pentium M.

Table 3
Improvement percentage of the number of times in which the correct orders are selected by the modified criteria (up, assuming d is known, down, assuming d is unknown)

M	$T=30$	BIC*	AIC*	AICc*	AICu*	C_1^*	Cc*	Cu*
d Known								
1	(p_1, p_2)	23.20	30.31	7.00	2.21	30.57	4.40	3.12
2	(p_1, p_2)	17.97	27.57	5.90	4.22	19.04	6.87	3.48
	$T=50$							
1	(p_1, p_2)	21.90	50.50	9.06	4.81	38.11	8.51	5.02
2	(p_1, p_2)	23.05	42.12	11.21	5.48	29.71	9.86	4.26
	$T=100$							
1	(p_1, p_2)	7.13	20.05	10.05	4.78	22.62	12.83	4.44
2	(p_1, p_2)	6.22	22.19	10.09	5.56	15.82	12.73	4.13
d Unknown								
	$T=30$							
1	(p_1, p_2, d)	38.51	41.02	11.66	7.40	28.34	6.39	3.67
2	(p_1, p_2, d)	22.68	43.87	5.20	2.08	25.12	2.73	−1.21
	$T=50$							
1	(p_1, p_2, d)	31.31	27.50	11.11	7.43	39.25	14.28	10.42
2	(p_1, p_2, d)	24.65	74.24	6.76	1.90	24.47	5.84	1.09
	$T=100$							
1	(p_1, p_2, d)	19.94	42.22	36.19	15.95	28.71	16.61	10.32
2	(p_1, p_2, d)	4.74	27.49	12.76	0.79	21.27	8.70	6.08

Acknowledgments

We would like to thank the associate editor and two anonymous referees for their helpful comments, and Jan De Gooijer for making his code available to us. We acknowledge financial support by Ministerio de Educación y Ciencia Grant SEJ2004-03303 and Comunidad de Madrid Grant HSE/0174/2004. The first author also acknowledges financial support by Xunta de Galicia under the Isidro Parga Pondal Program.

Appendix A.

Proof of Theorem 1. Shibata (1980) considers order selection criteria of the form:

$$S_T^{\mathrm{o}}(p) = (T - p_{\max} + \delta_T(p) + 2p)\widehat{\sigma}_p^2.$$

The order chosen for the selection criteria $S_T^{\mathrm{o}}(p)$ is efficient if $\delta_T(p)$ verifies the conditions imposed in Theorem 4.2 of Shibata (1980)

1. $\operatorname{p\,lim}_{T\to\infty}\left(\max_{1\leqslant p\leqslant p_{\max}}\dfrac{|\delta_T(p)|}{T-p_{\max}}\right)=0,$
2. $\operatorname{p\,lim}_{T\to\infty}\left(\max_{1\leqslant p\leqslant p_{\max}}\dfrac{|\delta_T(p)-\delta_T(p_T^*)|}{(T-p_{\max})L_T(p)}\right)=0,$

where p lim denotes limit in probability, $L_T(p)$, is the following function:

$$L_T(p) = \frac{p\sigma_a^2}{T-p_{\max}} + \sum_{i=p+1}^{\infty}\sum_{j=p+1}^{\infty}\phi_i\phi_j\Sigma_{ij},$$

where $\Sigma_{ij} = Cov(x_t, x_{t-|i-j|})$ and p_T^* is a sequence of positive integers with $1 \leqslant p_T^* \leqslant p_{\max}$ which attain the minimum of $L_T(p)$ for each T (see Shibata, 1980, p. 154). The AIC can be written in terms of $S_T^{\mathrm{o}}(p)$ taking $\delta_T(p) = \delta_T^{\mathrm{AIC}}(p) = T\exp(\frac{2p}{T}) - (T - p_{\max}) - 2p$. Shibata (1980) has shown that this term verifies the two conditions, and this gives the asymptotic efficiency of AIC. We can write AIC* in terms of $S_T^{\mathrm{o}}(p)$ taking $\delta_T(p) = \delta_T^{\mathrm{AIC}^*}(p) = T\exp(\frac{2p}{T})(\log|Q(\widehat{\beta}_p)|)^{1/T} - (T - p_{\max}) - 2p$. Therefore,

$$\delta_T^{\mathrm{AIC}^*}(p) = \delta_T^{\mathrm{AIC}}(p) - T\exp\left(\frac{2p}{T}\right)(1 - (\log|Q(\widehat{\beta}_p)|)^{1/T}).$$

We show that $\delta_T^{\mathrm{AIC}^*}(p)$ verifies both conditions. First we write,

$$\frac{|\delta_T^{\mathrm{AIC}^*}(p)|}{T - p_{\max}} = \left|\frac{\exp\left(\frac{2p}{T}\right)(\log|Q(\widehat{\beta}_p)|)^{1/T}}{1 - \frac{p_{\max}}{T}} - \frac{\frac{2p}{T}}{1 - \frac{p_{\max}}{T}} - 1\right|. \tag{17}$$

Hannan (1973) shows that $(\log|Q(\gamma)|)^{1/T} \to 1$, for every γ belonging to the parametric space, and consequently, $(\log|Q(\widehat{\beta}_p)|)^{1/T} \to 1$ and the limit when $T \to \infty$ of the maximum of the values (17) in the set $1 \leqslant p \leqslant p_{\max}$ is 0. This proves the first condition.

For the second condition, we write the following decomposition:

$$\begin{aligned}
&\frac{|\delta_T^{\mathrm{AIC}^*}(p) - \delta_T^{\mathrm{AIC}^*}(p_T^*)|}{(T - p_{\max})L_T(p)} \\
&\leqslant \frac{|\delta_T^{\mathrm{AIC}}(p) - \delta_T^{\mathrm{AIC}}(p_T^*)|}{(T - p_{\max})L_T(p)} \\
&\quad + \frac{\left|T\exp\left(\frac{2p_T^*}{T}\right)(1 - (\log|Q(\widehat{\beta}_{p_T^*})|)^{1/T}) - T\exp\left(\frac{2p}{T}\right)(1 - (\log|Q(\widehat{\beta}_p)|)^{1/T})\right|}{(T - p_{\max})L_T(p)}.
\end{aligned}$$

Shibata (1980) showed that the first term tends to 0 implying that AIC is efficient. For the second expression, for any p such that $1 \leqslant p \leqslant p_{\max}$ including p_T^*, it can be shown that,

$$\lim_{T\to\infty} T\exp\left(\frac{2p}{T}\right)(1 - (\log|Q(\widehat{\beta}_p)|)^{1/T}) = -\log\left(-\sum_{i=1}^{p} i\log(1 - \phi_{ii}^2(\beta_p))\right) < \infty.$$

As this limit is bounded for every p and $(T - p_{\max})L_T(p) \to \infty$ when $T \to \infty$, for every $1 \leqslant p \leqslant p_{\max}$, the second expression also tends to 0. Then, $\delta_T^{\mathrm{AIC}^*}(p)$ verifies the second condition. Therefore, AIC* is efficient. As AICc* is asymptotically equivalent to AIC*, AICc* is also efficient. □

Proof of Theorem 2. The BIC* can be written as follows by using one step ahead prediction variances:

$$\begin{aligned}
\mathrm{BIC}^*(p) &= T\log\widehat{\sigma}_p^2 + \sum_{t=1}^{T}\log v_t^2(p) + (p+1)\log T \\
&= T\log\widehat{\sigma}_p^2 + (p+1)\left(\frac{1}{p+1}\sum_{t=1}^{T}\log v_t^2(p) + \log T\right).
\end{aligned}$$

Now, note that the last term in the previous expression verifies:

$$\frac{1}{p+1}\sum_{t=1}^{T}\log v_t^2(p) + \log T \to \infty, \quad \frac{1}{T}\left(\frac{1}{p+1}\sum_{t=1}^{T}\log v_t^2(p) + \log T\right) \to 0.$$

This shows that the criterion BIC* is under the conditions of Theorem 3 in Hannan (1980, p. 1073), implying that BIC* is consistent. □

References

Akaike, H., 1969. Fitting autoregressive models for prediction. Ann. Inst. Statist. Math. 21, 243–247.

Akaike, H., 1973. Information theory and an extension of the maximum likelihood principle. In: Proceedings of the Second International Symposium on Information Theory. Akademiai Kiadó, Budapest, pp. 267–281.

Campbell, E.P., 2004. Bayesian selection of threshold autoregressive models. J. Time Ser. Anal. 25, 467–482.

Chan, K.S., 1993. Consistency and limiting distribution of the least squares estimator of a threshold autoregressive model. Ann. Statist. 21, 520–533.

De Gooijer, J.G., 2001. Cross-validation criteria for SETAR model selection. J. Time Ser. Anal. 22, 267–281.

Durbin, J., 1960. The fitting of time series models. Rev. Internat. Statist. Inst. 28, 233–244.

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

Hannan, E.J., 1980. Estimation of the order of an ARMA process. Ann. Statist. 8, 1071–1081.

Hannan, E.J., Quinn, B.G., 1979. The determination of the order of an autoregression. J. Roy. Statist. Soc. Ser. B 41, 190–195.

Harvey, A.C., 1981. Forecasting, Structural Time Series Models and the Kalman Filter. Cambridge University Press, Cambridge.

Hurvich, C.M., Tsai, C., 1989. Regression and time series model selection in small samples. Biometrika 76, 297–307.

Hurvich, C.M., Shumway, R., Tsai, C., 1990. Improved estimators of Kullback-Leibler information for autoregressive model selection in small samples. Biometrika 77, 709–719.

Kapetanios, G., 2001. Model selection in threshold models. J. Time Ser. Anal. 22, 733–754.

van der Leeuw, J., 1994. The covariance matrix of ARMA errors in closed form. J. Econometrics 63, 397–405.

McQuarrie, A., Shumway, R., Tsai, C.L., 1997. The model selection criterion AICu. Statist. Probab. Lett. 34, 285–292.

Öhrvik, J., Schoier, G., 2005. SETAR model selection—a bootstrap approach. Comput. Statist. 20, 559–573.

Peña, D., Rodriguez, J., 2006. The log of the determinant of the autocorrelation matrix for testing goodness of fit in time series. J. Statist. Plann. Inference 136, 2706–2718.

Ramsey, F.L., 1974. Characterization of the partial autocorrelation function. Ann. Statist. 2, 1296–1301.

Schwarz, G., 1978. Estimating the dimension of a model. Ann. Statist. 6, 461–464.

Shibata, R., 1980. Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. Ann. Statist. 8, 147–164.

Stoica, P., Eykhoff, P., Janssen, P., Söderström, T., 1986. Model-structure selection by cross-validation. Internat. J. Control 43, 1841–1878.

Tong, H., 1990. Non-linear Time Series: A Dynamical System Approach. Oxford University Press, Oxford.

Unnikrishnan, N.K., 2004. Bayesian subset model selection for time series. J. Time Ser. Anal. 25, 671–690.

Wong, C.S., Li, W.K., 1998. A note on the corrected Akaike information criterion for the threshold autoregressive models. J. Time Ser. Anal. 19, 113–124.