Available online at www.sciencedirect.com
ScienceDirect
Statistics & Probability Letters 77 (2007) 896–900
STATISTICS & PROBABILITY LETTERS
www.elsevier.com/locate/stapro

On the connection between model selection criteria and quadratic discrimination in ARMA 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 15 February 2006; received in revised form 13 October 2006; accepted 19 December 2006
Available online 19 January 2007

Abstract

This article establishes the connection between quadratic discrimination and model selection criterion in the ARMA framework. We show that analyzing model selection in ARMA time series models as a quadratic discrimination problem provides a unifying approach for deriving model selection criteria.
© 2007 Elsevier B.V. All rights reserved.

Keywords: Asymptotic efficiency; Consistency; Model selection criteria; Quadratic discrimination rule

1. Introduction

Most of the model selection criteria for linear time series can be written as $\min_k\{\log|\widehat{\Sigma}_k| + (k+1)\times C(T,k+1)\}$, where k is the number of estimated parameters for the mean function of the process, $\widehat{\Sigma}_k$ is the maximum likelihood estimation of the covariance matrix of the series $x=(x_1,\ldots,x_T)'$, T is the sample size and $C(T,k+1)$ is a function depending on T and $k+1$. These criteria can be classified into two groups. The first one includes the consistent criteria that, under the assumption that the data come from a finite order autoregressive moving average process, have a probability of obtaining the true order of the model that goes to one when the sample size increases. The Bayesian information criterion, BIC, by Schwarz (1978), where $C(T,k+1)=\log(T)$, and the Hannan and Quinn (1979) criterion, HQC, where $C(T,k+1)=2m\log\log(T)$ with $m>1$, are consistent criteria. The second group includes the efficient criteria, that select asymptotically the order which produces the least mean square prediction error. The final prediction error criterion, FPE, by Akaike (1969), where $C(T,k+1)=(T/(k+1))\log((T+k+1)/(T-(k+1)))$, the Akaike's information criterion, AIC, by Akaike (1973), where $C(T,k+1)=2$ and the corrected Akaike's information criterion, AICc, by Hurvich and Tsai (1989), where $C(T,k+1)=\frac{2T}{T-(k+1)-1}$, are efficient criteria. These criteria have been derived from different points of view. The BIC approach uses the posterior probabilities of the models. The HQC has been derived to be a consistent criterion such that $C(T,k+1)/T$ converges to 0 as fast as possible. The FPE selects the model that minimizes the one step ahead square

*Corresponding author. Tel.: +34 981 563100x13207; fax: +34 981 597054.
E-mail address: pgaleano@usc.es (P. Galeano).

0167-7152/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.spl.2006.12.010

prediction error. The AIC is an estimator of the expected Kullback–Leibler distance between the true and the fitted model. The AICc is a bias correction form of the AIC that appears to work better in small samples.

In this article we consider model selection as a discrimination problem and show that the AIC, AICc and BIC criteria can be derived as approximations to a quadratic discriminant rule, showing the connection between discrimination and model selection in linear Gaussian time series. The main contribution of this article is to view the model selection problem as a kind of discrimination analysis and present an unified approach of criteria proposed in the literature from different points of view. The technical details in both maximum likelihood and Bayesian points of view are included for completeness.

The rest of this paper is organized as follows. Section 2 briefly review the quadratic discriminant rule in ARMA time series. Sections 3 and 4 show the connection between discrimination and model selection criterion from a maximum likelihood and Bayesian approaches, respectively.

2. The quadratic discriminant rule for ARMA time series models

The discrimination problem in time series appears as follows. Suppose it is known that a given time series, $x = (x_1, \ldots, x_T)'$, has been generated by one of the models M_j, $j = 1, \ldots, j_{\max}$. From the Bayesian point of view we also know the prior probabilities $p(M_j)$. The objective is to select the data generating model given the time series data. We assume that the models M_j are causal and invertible Gaussian processes given by $x_t = \mu_{jt} + n_{jt}$, where μ_{jt} are deterministic mean functions and n_{jt} are zero mean ARMA models of the form $\phi_j(B)n_{jt} = \theta_j(B)a_{jt}$, where $\phi_j(B)$ and $\theta_j(B)$ are polynomials in the lag operator B such that $Bx_t = x_{t-1}$, with no common roots. The series a_{jt} are white noise innovations with variance σ_j^2. The simplest discriminant problem is to assume that the deterministic functions μ_{jt} are different, but the covariance matrices of x under each ARMA model n_{jt}, Σ_j, are all equal to Σ, which corresponds to the situation in which all the models have the same ARMA structure. Calling $\mu_j = (\mu_{j1}, \ldots, \mu_{jT})'$, this is equivalent to consider the hypothesis $M_j : x \in N_T(\mu_j, \Sigma)$, and we have that $p(x \mid M_j) = (2\pi)^{-T/2}|\Sigma|^{-1/2}\exp(-\frac{1}{2}(x-\mu_j)'\Sigma^{-1}(x-\mu_j))$, $j = 1, \ldots, j_{\max}$.

Maximizing the likelihood of the data implies minimizing the Mahalanobis distance between the data and the vector of marginal means. The same conclusion is obtained from the Bayesian point of view assuming equal prior probabilities $p(M_j) = 1/j_{\max}$ and maximizing the posterior probability of choosing the true model. A more interesting case appears when the ARMA models are different, that is, $M_j : x \in N_T(\mu_j, \Sigma_j)$, for $j = 1, \ldots, j_{\max}$. Then, the standard quadratic classification rule selects the model i if,

$$i = \underset{1 \leqslant j \leqslant j_{\max}}{\arg\max}(2\pi)^{-T/2}|\Sigma_j|^{-1/2}\exp\left(-\frac{1}{2}(x-\mu_j)'\Sigma_j^{-1}(x-\mu_j)\right) \tag{1}$$

and the Bayesian rule selects the model i if,

$$i = \underset{1 \leqslant j \leqslant j_{\max}}{\arg\max}\, p(M_j)(2\pi)^{-T/2}|\Sigma_j|^{-1/2}\exp\left(-\frac{1}{2}(x-\mu_j)'\Sigma_j^{-1}(x-\mu_j)\right). \tag{2}$$

In the next two sections the rules (1) and (2) are approximated in several ways and the AIC, AICc and BIC criteria are obtained when the data, $x = (x_1, \ldots, x_T)'$, have been generated by the class of ARMA Gaussian processes given by $x_t - \phi_1 x_{t-1} - \cdots - \phi_p x_{t-p} = a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q}$, $t = \ldots, -1, 0, 1, \ldots$, where a_t is a sequence of independent Gaussian distributed random variables with zero mean and variance $\sigma_{p,q}^2$ and we assume that $p \in \{0, \ldots, p_{\max}\}$ and $q \in \{0, \ldots, q_{\max}\}$, where $p_{\max}$ and $q_{\max}$ are some fixed upper bounds. We call the ARMA(p, q) model $M_{p,q}$, where $\beta_{p,q} = (\phi_{1p}, \ldots, \phi_{pp}, 0, \ldots, 0, \theta_{1q}, \ldots, \theta_{qq}, 0, \ldots, 0)'$ is a $(p_{\max} + q_{\max}) \times 1$ vector of parameters for the $M_{p,q}$ model and we define $\alpha_{p,q} = (\beta_{p,q}', \sigma_{p,q}^2)'$. We denote the parameters of the model that have generated the data as $\alpha_0 = (\beta_0', \sigma_0^2)'$. In this case, let $\widehat{\beta}_{p,q}$ and $\widehat{\sigma}_{p,q}^2$ be the maximum likelihood estimates of the vector of parameters $\beta_{p,q}$ and the innovations variance, respectively. The covariance matrix of x assuming the model $M_{p,q}$ can be written as $\Sigma_T(\alpha_{p,q}) = \sigma_{p,q}^2 Q_T(\beta_{p,q})$, where $Q_T(\beta_{p,q})$ is a $T \times T$ matrix

depending on the parameters $\beta_{p,q}$. Let $Q_T(\beta_{p,q}) = L(\beta_{p,q})L'(\beta_{p,q})$ be the Cholesky decomposition of $Q_T(\beta_{p,q})$. We denote, $a(\beta_{p,q}) = L(\beta_{p,q})^{-1}x$ and $S_x(\beta_{p,q}) = a(\beta_{p,q})'a(\beta_{p,q})$. We consider the following assumption:

Assumption 1. The models $M_{p,q}$ are causal, invertible and stationary and with polynomials $1 - \phi_1 B - \cdots - \phi_p B^p$ and $1 - \theta_1 B - \cdots - \theta_q B^q$ with no common roots.

3. A maximum likelihood approach

From (1), the discriminant rule assigns the data $x = (x_1, \ldots, x_T)'$, to the model $M_{p,q}$ with parameters $\alpha_{p,q}$ that maximizes $p(x \mid M_{p,q}) = p(x \mid \alpha_{p,q})$. In practice, the parameters are unknown and it is well known that if we substitute the unknown parameters, $\alpha_{p,q}$, by its maximum likelihood estimates, $\widehat{\alpha}_{p,q}$, maximizing the likelihood will always choose the model with the largest number of parameters. To avoid this problem, we need to obtain a suitable approximation of the quadratic rule. A first attempt to do that is to approximate $\log p(x \mid \alpha_{p,q})$ by

$$E_{\alpha_0}[\log p(y|\widehat{\alpha}_{p,q})] = \int \log p(y|\widehat{\alpha}_{p,q}) p(y|\alpha_0)\, \mathrm{d}y, \tag{3}$$

and select the model that maximizes (3), that is, the model that maximizes the expectation with respect to future observations generated by the true model, which has parameters α_0. Note that this rule selects the model which minimizes the Kullback–Leibler divergence to the true one. As,

$$E_{\alpha_0}\left[\log \frac{p(y|\alpha_0)}{p(y|\widehat{\alpha}_{p,q})}\right] = \int \log \frac{p(y|\alpha_0)}{p(y|\widehat{\alpha}_{p,q})} p(y|\alpha_0)\, \mathrm{d}y \geqslant 0$$

and the integral is always positive, minimizing it implies making $p(y|\widehat{\alpha}_{p,q})$ as close as possible to $p(y|\alpha_0)$, in the Kullback–Leibler divergence. This rule computes the log-likelihood of each model using the estimates $\widehat{\alpha}_{p,q}$ based on the sample and then compute the expectation with respect to future observations. The model chosen is the one which leads to a larger expected value of this maximized log-likelihood. Note that this approach takes into account the uncertainty about new observations but not the uncertainty in the parameter estimates. The following lemma shows that this simple approach fails to provide a suitable rule for selecting an ARMA model among the set of candidates.

Lemma 1. *Under Assumption* 1,

1. *if the parameters are evaluated at* $\widehat{\beta}_{p,q}$ *and* $(T/(T-(p+q)))\widehat{\sigma}^2_{p,q}$:

$$E_{\alpha_0}[\log p(y|\widehat{\alpha}_{p,q})] = -\frac{T}{2}(\log 2\pi + 1) - \frac{1}{2}\log|\Sigma_T(\widehat{\beta}_{p,q})| - (p+q+1) + \mathrm{O_p}(1), \tag{4}$$

2. *if the parameters are evaluated at* $\widehat{\beta}_{p,q}$ *and* $\widehat{\sigma}^2_{p,q}$:

$$E_{\alpha_0}[\log p(y|\widehat{\alpha}_{p,q})] = -\frac{T}{2}(\log 2\pi + 1) - \frac{1}{2}\log|\Sigma_T(\widehat{\beta}_{p,q})| - \frac{T(p+q+1)}{T-(p+q+1)-1} + \mathrm{O_p}(1). \tag{5}$$

Proof. From (1), we have that

$$E_{\alpha_0}[\log p(y|\widehat{\alpha}_{p,q})] = -\frac{T}{2}\log 2\pi - \frac{1}{2}\log|\Sigma_T(\widehat{\beta}_{p,q})| - \frac{1}{2}E_{\alpha_0}\left[\frac{S_y(\widehat{\beta}_{p,q})}{\widehat{\sigma}^2_{p,q}}\right],$$

where $S_y(\widehat{\beta}_{p,q}) = y'Q_T^{-1}(\widehat{\beta}_{p,q})y$. Assuming that $M_{p,q}$ is the model that actually generates $x = (x_1, \ldots, x_T)'$, Brockwell and Davis (1991) showed that

$$E_{\alpha_0}\left[\frac{S_y(\widehat{\beta}_{p,q})}{\widehat{\sigma}^2_{p,q}}\right] = \frac{T(T+p+q)}{(T-p-q-2)} + \mathrm{O_p}(1), \tag{6}$$

that gives (5). On the other hand, as $\log|\Sigma_T(\widehat{\beta}_{p,q})| = T\log\widehat{\sigma}^2_{p,q} + \log|Q_T(\widehat{\beta}_{p,q})|$ and $T\log(1-(p+q)/T) = -(p+q)+\mathrm{o}(1)$, we have that

$$\begin{aligned} T\log 2\pi + T\log\frac{T}{T-(p+q)}\widehat{\sigma}^2_{p,q} + \log|Q_T(\widehat{\beta}_{p,q})| &= T\log 2\pi - T\log\left(1-\frac{(p+q)}{T}\right) + T\log\widehat{\sigma}^2_{p,q} \\ &\quad + \log|Q_T(\widehat{\beta}_{p,q})| \\ &= T\log 2\pi + T\log\widehat{\sigma}^2_{p,q} + (p+q) + \log|Q_T(\widehat{\beta}_{p,q})| \\ &\quad + \mathrm{o_p}(1). \end{aligned}$$

From (6),

$$E_{\alpha_0}\left[\frac{S_y(\widehat{\beta}_{p,q})}{\dfrac{T}{T-(p+q)}\widehat{\sigma}^2_{p,q}}\right] = (T+p+q) + \frac{2(T+p+q)}{(T-p-q-2)} + \mathrm{O_p}(1) = T+p+q+2+\mathrm{O_p}(1),$$

which proves (4). □

This lemma shows that (4) and (5) include terms that are $\mathrm{O_p}(1)$ which are of the same order as the penalty terms. Following Brockwell and Davis (1991), the $\mathrm{O_p}(1)$ remainder term reduces to a component o(1) and a component which has expectation zero. Thus, we see that we cannot avoid taking into account the uncertainty about the parameter estimates. We can solve this problem by taking also the expectation with respect to the distribution of the estimate, $\widehat{\alpha}_{p,q}$. Then, we select the model which leads to a larger value of:

$$E_{\widehat{\alpha}_{p,q}}[E_{\alpha_0}[\log p(y|\widehat{\alpha}_{p,q})]] = \int\int \log p(y|\widehat{\alpha}_{p,q})p(y|\alpha_0)f(\widehat{\alpha}_{p,q}|\alpha_0)\,\mathrm{d}y\,\mathrm{d}\widehat{\alpha}_{p,q},$$

where $f(\widehat{\alpha}_{p,q}|\alpha_{p_0,q_0})$ is the distribution of the estimate and $\widehat{\alpha}_{p,q}$ and y are assumed to be independent. Thus, the rule selects the model that maximizes the expected value with respect to the two sources of uncertainty: the distribution of future observations and the distribution of the estimate. Note that this is equivalent to the criterion proposed by Akaike (1969, 1973) from different arguments, and therefore, after taking expectations in the expression (4), we get the criterion:

$$\mathrm{AIC}(p,q) = \log|\Sigma_T(\widehat{\beta}_{p,q})| + 2(p+q+1) \tag{7}$$

while (5) leads to the criterion:

$$\mathrm{AICc}(p,q) = \log|\Sigma_T(\widehat{\beta}_{p,q})| + \frac{2T(p+q+1)}{T-(p+q+1)-1}, \tag{8}$$

which are the expression of both criteria, as given in Hurvich et al. (1990).

4. A Bayesian approach

We analyze the rule in (2) taking into account that this approach requires prior probabilities of the models, $p(M_{p,q})$ and the parameters, $p(\alpha_{p,q}|M_{p,q})$. The Bayesian point of view of maximizing the posterior probability has been extensively considered, see Schwarz (1978), Chow (1981), Haughton (1988) or Raftery et al. (1996). Note that when computing this posterior probability we automatically take into account the two sources of uncertainty discussed in the previous section.

Lemma 2. *Under Assumption* 1,

$$\begin{aligned} \log p(x|M_{p,q}) &= \tfrac{1}{2}(p+q+1-T)\log(2\pi) - \tfrac{1}{2}(p+q+1)\log(T) - \tfrac{1}{2}\log|\Sigma_T(\widehat{\beta}_{p,q})| \\ &\quad - \tfrac{1}{2}T + \log p(\widehat{\alpha}_{p,q}|M_{p,q}) + \mathrm{O_p}(1). \end{aligned} \tag{9}$$

Proof. Let, $h(\alpha_{p,q}) = -(T/2)\log(2\pi) - \frac{1}{2}\log|\Sigma_T(\alpha_{p,q})| - \frac{1}{2}\, x'\Sigma_T(\alpha_{p,q})^{-1}x + \log p(\alpha_{p,q}|M_{p,q})$. Then, applying the Laplace's method, see Tierney and Kadane (1986), we obtain

$$p(x|M_{p,q}) \simeq (2\pi)^{(p+q+1-T)/2}|H(\widehat{\alpha}_{p,q})|^{1/2}|\Sigma_T(\widehat{\alpha}_{p,q})|^{-1/2}\exp(-\tfrac{1}{2}x'\Sigma_T(\widehat{\alpha}_{p,q})^{-1}x)p(\widehat{\alpha}_{p,q}|M_{p,q}),$$

where $H(\widehat{\alpha}_{p,q})$ is minus the inverse Hessian of h evaluated at $\widehat{\alpha}_{p,q}$. The inverse of the observed information matrix is asymptotically equal to T times a constant matrix (see, for instance, Raftery et al., 1996), so that, $\log|H(\widehat{\alpha}_{p,q})| = -(p+q+1)\log T + \mathrm{O_p}(1)$, and

$$\log p(x|M_{p,q}) = \tfrac{1}{2}(p+q+1-T)\log(2\pi) - \tfrac{1}{2}(p+q+1)\log T - \tfrac{1}{2}(\log|\Sigma_T(\widehat{\alpha}_{p,q})| + T) + \log p(\widehat{\alpha}_{p,q}|M_{p,q}) + \mathrm{O_p}(1),$$

which proves the stated result. □

Taking the same prior probabilities for all the parameters and ignoring constant terms, (9) leads to the criterion

$$\mathrm{BIC}(p,q) = \log|\Sigma_T(\widehat{\beta}_{p,q})| + \log(T)(p+q+1). \tag{10}$$

The criteria (7), (8) and (10) can be written as

$$\min_{(p,q)}\{\log|\Sigma_T(\widehat{\beta}_{p,q})| + (p+q+1)\times C(T,p+q+1)\}, \tag{11}$$

where the term $|\Sigma_T(\widehat{\beta}_{p,q})|$ is easily obtained from the maximized log-likelihood, $\log p(x|\widehat{\alpha}_{p,q})$, due that $\log|\Sigma(\widehat{\alpha}_{p,q})| = -2\log p(x|\widehat{\alpha}_{p,q}) - T(\log(2\pi)+1)$.

Acknowledgments

We acknowledge financial support by MEC Grant SEJ2004-03303 and CAM Grant HSE/0174/2004. The first author acknowledges financial support by Xunta de Galicia under the Isidro Parga Pondal Program.

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.

Brockwell, P.J., Davis, R.A., 1991. Time Series: Theory and Methods, second ed. Springer, New York.

Chow, G.C., 1981. A comparison of the information and posterior probability criteria for model selection. J. Econometrics 16, 21–33.

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

Haughton, D.M.A., 1988. On the choice of a model to fit data from an exponential family. Ann. Statist. 16, 342–355.

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.

Raftery, A.E., Madigan, D., Volinsky, C.T., 1996. Accounting for model uncertainty in survival analysis improves predictive. In: Bernardo, J.M., Berger, J.O., Dawid, A.P., Smith, A.F.M. (Eds.), Bayesian Statistics, vol. 5. Oxford University Press, Oxford, pp. 323–349.

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

Tierney, L., Kadane, J.B., 1986. Accurate approximations for posterior moments and marginal densities. J. Amer. Statist. Assoc. 81, 82–86.