Model selection criteria [*]

Jean-Marie Dufour [†]
Université de Montréal

First version: March 1991
Revised: July 1998
This version: April 7, 2002
Compiled: April 7, 2002, 4:10pm

---

[*] This work was supported by the Canada Research Chair Program (Chair in Econometrics, Université de Montréal), the Canadian Network of Centres of Excellence [program on *Mathematics of Information Technology and Complex Systems* (MITACS)], the Canada Council for the Arts (Killam Fellowship), the Natural Sciences and Engineering Research Council of Canada, the Social Sciences and Humanities Research Council of Canada, and the Fonds FCAR (Government of Québec).

[†] Canada Research Chair Holder (Econometrics). Centre interuniversitaire de recherche en analyse des organisations (CIRANO), Centre de recherche et développement en économique (C.R.D.E.), and Département de sciences économiques, Université de Montréal. Mailing address: Département de sciences économiques, Université de Montréal, C.P. 6128 succursale Centre-ville, Montréal, Québec, Canada H3C 3J7. TEL: 1 514 343 2400; FAX: 1 514 343 5831; e-mail: jean.marie.dufour@umontreal.ca. Web page: http://www.fas.umontreal.ca/SCECO/Dufour .

Contents

1. Introduction

On using usual preliminary specification and residual-based diagnostics, several models often appear to be essentially equivalent for representing the behavior of a time series. In such cases, it can be quite useful to use model selection criteria.

Suppose X_t follows an $ARIMA\,(p, d, q)$ process:

$$\varphi_p\,(B)\,(1-B)^d\,X_t = \bar{\mu} + \theta_q\,(B)\,u_t\ ,\ t \geq 1-d$$

where $\{u_t : t \in Z\} \sim BB\,(0, \sigma^2)\,.$

This model is estimated from the series differentiated d times: $W_t = (1-B)^d\,X_t,\ t = 1,\ ...,\ T$. Let

$$\hat{\sigma}_W^2 = \sum_{t=1}^{T} \left(W_t - \bar{W}\right)^2 / T$$

where $\bar{W} = \sum_{t=1}^{T} W_T / T$, the sample variance of W_t, and let $\hat{\sigma}_T^2$ the maximum likelihood (ML) estimator of σ^2 :

$$\hat{\sigma}_T^2 = \sum \hat{u}_t^2 / T.$$

2. Predictive performance criteria

Since σ^2 is the variance of the one-step ahead error prediction error, it is natural to

a) minimize $\hat{\sigma}_T^2$,

or

b) maximize $R^2 = 1 - (\hat{\sigma}_T^2\,/\,\hat{\sigma}_W^2)$.

These two criteria are equivalent. However, $\hat{\sigma}_T^2$ automatically decreases when p or q increases. In order to penalize models which contain too many parameters, it is preferable to use statistics which involve a correction for the number of degrees of freedom:

c) minimize

$$s_T^2 = \frac{T}{T-p-q}\hat{\sigma}_T^2 = \sum_t \hat{u}_t^2 / \left(T-p-q\right)$$

or

d) maximize

$$\bar{R}^2 = 1 - \frac{s_T^2}{s_W^2} = 1 - \frac{T-1}{T-p-q}\frac{\hat{\sigma}_T^2}{\hat{\sigma}_W^2}$$

where $s_W^2 = \sum_{t=1}^{T} \left(W_t - \bar{W}\right)^2 / \left(T-1\right).$

3. Information criteria

Another approach consists in evaluating the "distance" between the selected model and the true (unknown) model. Let $f\left(W\right)$ the density associated with the postulated model and $f_o\left(W\right)$ the density of the true model, where $W = \left(W_1,\, \ldots\, ,\, W_T\right)'$. One such distance consists in using the *Kullback distance*:

$$\begin{aligned} I\left(f, f_o\right) &= \int \log\left[f_o\left(w\right)/f\left(w\right)\right] f_o\left(w\right) dw \\ &= \underset{f_o}{E}\left\{\log\left[f_o\left(W\right)/f\left(W\right)\right]\right\} \\ &= \underset{f_o}{E}\left\{\log\left[f_o\left(W\right)\right]\right\} - \underset{f_o}{E}\left\{\log\left[f\left(W\right)\right]\right\} \ . \end{aligned}$$

Minimizing $I\left(f, f_o\right)$ with respect to f is equivalent to minimizing $-E\left\{\log\left[f\left(W\right)\right]\right\}$. We obtain an information criterion by selecting an "estimator" of $-\underset{f_o}{E}\left\{\log\left(f\right)\right\}$. These different criteria take the following general form (up to an additive constant):

$$IC^* = -\frac{1}{T}\log\left(f\right) + \alpha\left(T\right)\left(p+q\right)$$

where $\alpha\left(T\right)$ is a decreasing function of T. We then try to minimize IC^*.

In the case where f is a normal density, IC^* takes the equivalent form:

$$IC = \log\left(\hat{\sigma}_T^2\right) + \alpha\left(T\right)\left(p+q\right) \ .$$

Different criteria are obtained by selecting different functions $\alpha\left(T\right)$. The most important ones are:

a) $\alpha\left(T\right) = 2/T$ [Akaike (1969)];

b) $\alpha\left(T\right) = \log\left(T\right)/T$ [Schwarz (1978)];

c) $\alpha\left(T\right) = c\log\left[\log\left(T\right)\right]/T$ where $c > 2$ [Hannan and Quinn (1979)].

The following criteria are then obtained:

a) Akaike criterion [Akaike (1969)]:

$$AIC\left(p, q\right) = \log\left(\hat{\sigma}_T^2\right) + \frac{2\left(p+q\right)}{T} \ ;$$

b) Schwarz criterion [Schwarz (1978)]:

$$BIC\left(p, q\right) = \log\left(\hat{\sigma}_T^2\right) + \left(p+q\right)\frac{\log\left(T\right)}{T} \ ;$$

c) Hannan-Quinn criterion [Hannan and Quinn (1979)]:

$$\varphi(p,q) = \log\left(\hat{\sigma}_T^2\right) + c\,(p+q)\,\frac{\log\left[\log(T)\right]}{T}\ , \text{ where } c > 2\ .$$

If we assume that the true values p_o and q_o satisfy $0 \leq p_o \leq P$ and $0 \leq q_o \leq Q$, and we minimize the information criterion over all the pairs $\{(p,q) : 0 \leq p \leq P\ ,\ 0 \leq q \leq Q\}$, it is possible to show [see Shibata (1976, 1980), Taniguchi (1980), Hannan and Quinn (1979), Hannan and Rissanen (1982)] that:

1. the Akaike criterion tends to identify values of p and q which are too large, i.e., the values of p and q that minimize AIC converge (as $T \to \infty$) towards values which are larger then p_o and q_o;
2. the values of p and q that minimize BIC converge towards p_o and q_o.

4. Bibliographic notes

For a general review of this topic, see Choi (1992, Chapter 4). For further discussion, see Brockwell and Davis (1991, Section 9.3), Lütkepohl (1991, Chapter 11) and Gouriéroux and Monfort (1997, Section 6.3). On the case of integrated series, see Paulsen (1984) and Toda and Yamamoto (1995).

References

Akaike, H. (1969), ‘Fitting autoregressive models for prediction’, *Annals of the Institute of Statistical Mathematics* **21**, 243–247.

Brockwell, P. J. and Davis, R. A. (1991), *Time Series: Theory and Methods*, second edn, Springer-Verlag, New York.

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

Gouriéroux, C. and Monfort, A. (1997), *Time Series and Dynamic Models*, Cambridge University Press, Cambridge, U.K.

Hannan, E. J. and Quinn, B. (1979), ‘The determination of the order of an autoregression’, *Journal of the Royal Statistical Society, Series B* **41**, 190–191.

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

Lütkepohl, H. (1991), *Introduction to Multiple Time Series Analysis*, Springer-Verlag, Berlin.

Paulsen, J. (1984), ‘Order determination of multivariate autoregressive time series with unit roots’, *Journal of Time Series Analysis* **5**, 115–127.

Schwarz, G. (1978), ‘Estimating the dimension of a model’, *The Annals of Statistics* **6**, 461–464.

Shibata, R. (1976), ‘Selection of the order of an autoregressive model by Akaike’s information criterion’, *Biometrika* **71**, 117–126.

Shibata, R. (1980), ‘Asymptotically efficient selection of the order of the model for estimating parameters of a linear process’, *The Annals of Statistics* **8**, 147–164.

Taniguchi, M. (1980), ‘On the selection of the order of the spectral density of a stationary process’, *Annals of the Institute of Statistical Mathematics* **32**, 401–409.

Toda, H. Y. and Yamamoto, T. (1995), ‘Statistical inference in vector autoregressions with possibly integrated processes’, *Journal of Econometrics* **66**, 225–250.