Information Criteria and Model Selection

Herman J. Bierens

Pennsylvania State University

March 12, 2006

1. Introduction

Let $L_n(k)$ be the maximum likelihood of a model with k parameters based on a sample of size n, and let k_0 be the correct number of parameters. Suppose that for $k > k_0$ the model with k parameters is nested in the model with k_0 parameters, so that $L_n(k_0)$ is obtained by setting $k - k_0$ parameters in the larger model to constants. Without loss of generality we may assume that these constants are zeros. Thus, denoting the likelihood function of the least parsimonious model by $\hat{L}_n(\theta)$, $\theta \in \Theta \subset \mathbb{R}^m$,

$$L_n(k) = \max_{\theta \in \Theta_k} \hat{L}_n(\theta), \text{ where } \Theta_k = \left\{ \theta = \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} \in \Theta \colon \theta_2 = 0 \in \mathbb{R}^{m-k} \right\} \tag{1}$$

for $k \le m$. Thus, the models with $k < k_0$ parameters are misspecified, and the models with $k > k_0$ parameters are correctly specified but over-parametrized.

The Akaike (1974, 1976), Hannan-Quinn (1979), and Schwarz (1978) information criteria for selecting the most parsimonious correct model are

Akaike: $c_n(k) = -2.\ln(L_n(k))/n + 2k/n$,

Hannan-Quinn: $c_n(k) = -2.\ln(L_n(k))/n + 2k.\ln(\ln(n))/n$,

Schwarz: $c_n(k) = -2.\ln(L_n(k))/n + k.\ln(n)/n$,

respectively. Since the Schwarz information criterion is derived using Bayesian arguments, this criterion is also known as the Bayesian Information Criterion (BIC).

These criteria take the general form

$$c_n(k) = -2.\ln(L_n(k))/n + k.\varphi(n)/n, \tag{2}$$

where $\varphi(n) = 2$ in the Akaike case, $\varphi(n) = 2.\ln(\ln(n))$ in the Hannan-Quinn case, and $\varphi(n) = \ln(n)$ in the Schwarz case. Using these criteria, the model is selected that corresponds to

$$\hat{k} = \text{argmin}_{k \le m} c_n(k). \tag{3}$$

2. Consistency

If $k < k_0$ then the model with k parameters is misspecified, so that

$$\text{plim}_{n\to\infty}\ln(L_n(k))/n < \text{plim}_{n\to\infty}\ln(L_n(k_0))/n. \tag{4}$$

Hence, it follows from (2), (4) and $\lim_{n\to\infty}\varphi(n)/n = 0$ that in all three cases

$$\begin{aligned}&\lim_{n\to\infty}P[c_n(k_0) \geq c_n(k)]\\ &= \lim_{n\to\infty}P[-2.\ln(L_n(k_0))/n + k_0.\varphi(n)/n \geq -2.\ln(L_n(k))/n + k.\varphi(n)/n]\\ &= \lim_{n\to\infty}P[\ln(L_n(k_0))/n - \ln(L_n(k))/n \leq 0.5(k_0-k).\varphi(n)/n] = 0,\end{aligned} \tag{5}$$

so that

$$\begin{aligned}\lim_{n\to\infty}P[\hat{k} < k_0] &\leq \lim_{n\to\infty}P[c_n(k_0) \geq c_n(k) \text{ for some } k < k_0]\\ &\leq \Sigma_{k<k_0}\lim_{n\to\infty}P[c_n(k_0) \geq c_n(k)] = 0\end{aligned} \tag{6}$$

For $k > k_0$ it follows from the likelihood ratio test that

$$2\big(\ln(L_n(k)) - \ln(L_n(k_0))\big) \to_d X_{k-k_0} \sim \chi^2_{k-k_0}, \tag{7}$$

where $\to_d$ indicates convergence in distribution. Then in the Akaike case,

$$n\big(c_n(k_0) - c_n(k)\big) = 2\big(\ln(L_n(k)) - \ln(L_n(k_0))\big) - 2(k-k_0) \to_d X_{k-k_0} - 2(k-k_0),$$

hence

$$\lim_{n\to\infty}P[c_n(k_0) > c_n(k)] = P[X_{k-k_0} > 2(k-k_0)] > 0.$$

Therefore, the Akaike criterion may asymptotically overshoot the correct number of parameters:

$$\lim_{n\to\infty}P[\hat{k} \geq k_0] = 1, \ \textit{but} \ \lim_{n\to\infty}P[\hat{k} > k_0] > 0,$$

Since in the Hannan-Quinn and Schwarz cases, $\lim_{n\to\infty}\varphi(n) = \infty$, (7) implies that in these two cases

$$\text{plim}_{n\to\infty} -2\big(\ln(L_n(k_0)) - \ln(L_n(k))\big)/\varphi(n) = 0$$

hence

$$\text{plim}_{n\to\infty}n(c_n(k_0) - c_n(k))/\varphi(n) = \text{plim}_{n\to\infty}-2\big(\ln(L_n(k_0)) - \ln(L_n(k))\big)/\varphi(n) + k_0-k = k_0-k \leq -1$$

so that

$$\lim_{n\to\infty}P[c_n(k_0) \geq c_n(k)] = 0.$$

This implies, similar to (6), that $\lim_{n\to\infty}P[\hat{k} > k_0] = 0.$ Thus, in the Hannan-Quinn and Schwarz cases,

$$\lim_{n\to\infty} P[\hat{k} = k_0] = 1. \tag{8}$$

Note that the consistency result (8) holds for any criterion of the type (2) with

$$\lim_{n\to\infty}\varphi(n)/n = 0 \text{ and } \lim_{n\to\infty}\varphi(n) = \infty, \tag{9}$$

for example, let $\varphi(n) = \sqrt{n}$.

3. Applications

3.1 VAR and AR model selection

Let $L_n(k)$ be the maximum likelihood of a d-variate Gaussian VAR(p) model,

$$Y_t = a_0 + \Sigma_{j=1}^{p} A_j Y_{t-j} + U_t,\ U_t \sim i.i.d.\ N_d[0,\Sigma],$$

where $Y_t \in \mathbb{R}^d$ is observed for $t = 1-p,....,n$. Then $k = d + d^2.p$ and

$$\ln(L_n(k)) = -\tfrac{1}{2}n.d - \tfrac{1}{2}n.d.\ln(2\pi) - \tfrac{1}{2}n.\ln\left(\det(\hat{\Sigma}_p)\right),$$

where $\hat{\Sigma}_p$ is the maximum likelihood estimator of the error variance Σ. Hence,

$$-2.\ln(L_n(k))/n = \ln\left(\det(\hat{\Sigma}_p)\right) + d.\left(1 + \ln(2\pi)\right). \tag{10}$$

The second term does not depend on p. Therefore, the model is selected that corresponds to $\hat{p} = \text{argmin}_p c_n^{VAR}(p)$, where

Akaike: $c_n^{VAR}(p) = \ln(\det(\hat{\Sigma}_p)) + 2(d+d^2p)/n,$

Hannan-Quinn: $c_n^{VAR}(p) = \ln(\det(\hat{\Sigma}_p)) + 2(d+d^2p)\ln(\ln(n))/n,$

Schwarz: $c_n^{VAR}(p) = \ln(\det(\hat{\Sigma}_p)) + (d+d^2p)\ln(n)/n.$

Similarly, these criteria can also be used to determine the order p of an AR(p) model:

$$Y_t = \alpha_0 + \Sigma_{j=1}^{p}\alpha_j Y_{t-j} + U_t,\ U_t \sim i.i.d.\ N[0,\sigma^2],$$

where again $Y_t \in \mathbb{R}$ is observed for $t = 1-p,....,n$, simply by replacing d with 1 and $\det(\hat{\Sigma}_p)$ with the ML estimator $\hat{\sigma}_p^2$ of the error variance σ^2:

Akaike: $c_n^{AR}(p) = \ln(\hat{\sigma}_p^2) + 2(1+p)/n,$

Hannan-Quinn: $c_n^{AR}(p) = \ln(\hat{\sigma}_p^2) + 2(1+p)\ln(\ln(n))/n,$

Schwarz: $c_n^{AR}(p) = \ln(\hat{\sigma}_p^2) + (1+p)\ln(n)/n.$

3.2 ARMA model specification

Similarly, in the ARMA(p,q) case

$$Y_t = \alpha_0 + \Sigma_{j=1}^{p}\alpha_j Y_{t-j} + U_t - \Sigma_{j=1}^{q}\beta_j U_{t-j},\ U_t \sim i.i.d.\ N[0,\sigma^2],$$

these criteria become

$$\text{Akaike:} \quad c_n^{ARMA}(p,q) = \ln(\hat{\sigma}^2_{p,q}) + 2(1+p+q)/n,$$

$$\text{Hannan-Quinn:} \quad c_n^{ARMA}(p,q) = \ln(\hat{\sigma}^2_{p,q}) + 2(1+p+q)\ln(\ln(n))/n,$$

$$\text{Schwarz:} \quad c_n^{ARMA}(p,q) = \ln(\hat{\sigma}^2_{p,q}) + (1+p+q)\ln(n)/n,$$

where now $\hat{\sigma}^2_{p,q}$ is the ML estimator of the error variance σ^2 and n is the number of observations used in the ML estimation.

It can be shown [see Hannan (1980)] that in the case of common roots in the AR and MA polynomials the Hannan-Quinn and Schwarz criteria still select the correct orders p and q consistently: Given upper bounds $\bar{p} \geq p_0$ and $\bar{q} \geq q_0$, where p_0 and q_0 are the correct orders of an ARMA(p,q) process, we have $\lim_{n\to\infty} P[\hat{p} = p_0, \hat{q} = q_0] = 1$, where

$$(\hat{p},\hat{q}) = \text{argmin}_{0\leq p\leq\bar{p},0\leq q\leq\bar{q}}\, c_n^{ARMA}(p,q).$$

3.3 ARCH and GARCH models

If a model is extended to include ARCH or GARCH errors, it is recommended to subtract the term $1 + \ln(2\pi)$ from $-2.\ln(L_n(k))/n$ [see (10)] in the formula for the information criteria, in order to make these criteria comparable with those for the model without (G)ARCH errors. Thus,

$$\text{Akaike:} \quad c_n^{(G)ARCH}(k) = -2.\ln(L_n(k))/n + 2k/n - 1 - \ln(2\pi),$$

$$\text{Hannan-Quinn:} \quad c_n^{(G)ARCH}(k) = -2.\ln(L_n(k))/n + 2k.\ln(\ln(n))/n - 1 - \ln(2\pi),$$

$$\text{Schwarz:} \quad c_n^{(G)ARCH}(k) = -2.\ln(L_n(k))/n + k.\ln(n)/n - 1 - \ln(2\pi),$$

where again k is the number of parameters, including the (G)ARCH parameters.

References

Akaike, H. (1974): "A New Look at the Statistical Model Identification," *I.E.E.E. Transactions on Automatic Control*, *AC* 19, 716-723.

Akaike, H. (1976): "Canonical Correlation Analysis of Time Series and the Use of an Information Criterion," in R. K. Mehra and D. G. Lainotis (eds.), *System Identification: Advances and Case Studies*, Academic Press, New York, 52-107.

Hannan, E. J. (1980): "The Estimation of the Order of an ARMA Process", *Annals of Statistics*, 8, 1071-1081.

Hannan, E. J., and B. G. Quinn (1979): "The Determination of the Order of an Autoregression," *Journal of the Royal Statistical Society, B*, 41, 190-195.

Schwarz, G. (1978): "Estimating the Dimension of a Model," *Annals of Statistics*, 6, 461-464.