Information Criteria and Model Selection

Herman J. Bierens

Pennsylvania State University

August 27, 2004

1. Introduction

Let $L_n(k)$ be the 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. The Akaike, Hannan-Quinn, and Schwarz information criteria for selecting the number of parameters are

$$\begin{aligned}
&\text{Akaike:} && c_n(k) = -2.\ln(L_n(k))/n + 2k/n,\\
&\text{Hannan-Quinn:} && c_n(k) = -2.\ln(L_n(k))/n + 2k.\ln(\ln(n))/n,\\
&\text{Schwarz:} && c_n(k) = -2.\ln(L_n(k))/n + k.\ln(n)/n,
\end{aligned}$$

respectively, i.e., k_0 can be estimated by

$$\hat{k} = \text{argmin}_k c_n(k).$$

2. Consistency

If $k < k_0$ then $\text{plim}_{n\to\infty}\ln(L_n(k))/n < \text{plim}_{n\to\infty}\ln(L_n(k_0))/n$, hence in all three cases

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

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

$$2\big(\ln(L_n(k)) - \ln(L_n(k_0))\big) \to_d \chi^2_{k-k_0}, \qquad \textbf{(1)}$$

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

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

where $X_{k-k_0} \sim \chi^2_{k-k_0}$, 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.

Since **(1)** implies

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

and

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

it follows that in the Hannan-Quinn case,

$$\operatorname{plim}_{n\to\infty} n(c_n(k_0) - c_n(k))/\ln(\ln(n)) = 2(k-k_0) \geq 2$$

and in the Schwarz case,

$$\operatorname{plim}_{n\to\infty} n(c_n(k_0) - c_n(k))/\ln(n) = k-k_0 \geq 1,$$

so that in both cases

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

Consequently, in the Akaike case we have

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

whereas in the Hannan-Quinn and Schwarz cases,

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

If the model with $k > k_0$ is **not** nested in the correct model with k_0 parameters, then $L_n(k)$ is the quasi-likelihood of a misspecified model. Hence, similarly to the case $k < k_0$ we then have $\operatorname{plim}_{n\to\infty} \ln(L_n(k))/n < \operatorname{plim}_{n\to\infty} \ln(L_n(k_0))/n$, so that all three information criteria will asymptotically detect this case. Consequently, these information criteria can also be used for model selection in the

case that the explanatory variables do not have a natural ordering, but again only the Hannan-Quinn and Schwarz criteria will asymptotically yield the correct model.

3. Applications

3.1 VAR and AR model selection

If $L_n(k)$ is the likelihood of a Gaussian VAR(p) model,

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

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

$$\ln(L_n(k)) = -\tfrac{1}{2}n.m - \tfrac{1}{2}n.\ln[\det(\hat{\Sigma}_p)],$$

where $\hat{\Sigma}_p$ is the maximum likelihood estimator of the error variance Σ. Then we may use these criteria to determine the order p of the VAR:

$$\hat{p} = \operatorname{argmin}_p c_n^{VAR}(p),$$

where

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

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

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

These criteria can also be used to determine the order 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 m 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

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

these criteria become

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

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

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

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

4. Designing your own model selection criterion

Consider a collection of M models indexed by $m = 1,....,M$, where each model is characterized by a parameter vector θ_m of dimension k_m in a parameter space $\Theta_m \subset \mathbb{R}^{k_m}$. A correct model, correct in the sense that it has all the properties that it supposed to have, should be among these models. The parameters in each model are estimated by minimizing an objective function $Q_{m,n}(\theta_m)$, where n is the number of observations. Thus, each θ_m is estimated by

$$\theta_{m,n} = \underset{\theta\in\Theta_m}{\operatorname{argmin}} Q_{m,n}(\theta).$$

It is possible to set forth regularity conditions such that $\operatorname{plim}_{n\to\infty} Q_{m,n}(\theta_m) = \overline{Q}_m(\theta_m)$ exists and is finite, and denote

$$\theta_m^0 = \underset{\theta\in\Theta_m}{\operatorname{argmin}} \overline{Q}_m(\theta).$$

This is the pseudo-true value of the parameter vector of model m. Moreover, it is possible to set forth

further regularity conditions such that

$$\text{plim}_{n\to\infty} Q_{m,n}(\theta_{m,n}) = \overline{Q}_m(\theta_m^0).$$

Now divide this collection of models into two groups. Let the first group contain all the models that are correctly specified but possibly over-parametrized, say the models $m = 1,...,K < M$. Without loss of generality we may assume that model $m = 1$ is the most parsimonious of the models in this group, i.e., $k_1 < k_m$ for $m = 2,...,K$, and that model 1 is nested in the other models in this group, so that

$$\overline{Q}_1(\theta_1^0) = \overline{Q}_m(\theta_m^0) \text{ for } m = 2,...,K.$$

The other group of models, for $m = K+1,...,M$, consists of models that are misspecified, so that

$$\overline{Q}_1(\theta_1^0) < \overline{Q}_m(\theta_m^0) \text{ for } m = K+1,...,M.$$

The customized model selection criterion then takes the form

$$c_n(m) = Q_{m,n}(\theta_{m,n}) + \rho(k_m, n)$$

where $\rho(k_m, n)$ is a penalty function satisfying $\lim_{n\to\infty}\rho(k_m, n) = 0$ for $m = 1,...,M$. For $m = K+1,...,M$, we have $\text{plim}_{n\to\infty}(c_n(m) - c_n(1)) = \overline{Q}_m(\theta_m^0) - \overline{Q}_1(\theta_1^0) > 0$ and thus

$$\lim_{n\to\infty} P[c_n(m) > c_n(1)] = 1 \text{ for } m = K+1,...,M.$$

In order to select the most parsimonious model from the set of (over-parametrized) correct models, we need a further condition on the penalty $\rho(k_m, n)$. This further condition depends on the rate β_n of convergence in distribution of $\beta_n(Q_{1,n}(\theta_{1,n}) - Q_{m,n}(\theta_{m,n}))$ for $m = 2,...,K$. In the log-likelihood case $\beta_n = n$, , i.e., if $-n.Q_{m,n}(\theta_{m,n})$ is two-times a log-likelihood then we have seen above that $n[Q_{1,n}(\theta_{1,n}) - Q_{m,n}(\theta_{m,n})] \to_d \chi^2_{k_m - k_1}$.

Let α_n be a sequence of positive numbers converging to infinity such that $\alpha_n/\beta_n \to 0$. Then

$$\text{plim}_{n\to\infty}\alpha_n[Q_{1,n}(\theta_{1,n}) - Q_{m,n}(\theta_{m,n})] = 0 \text{ for } m = 2,...,K.$$

Next, specify the penalty function $\rho(k_m, n)$ such that

$$\lim_{n\to\infty}\alpha_n\rho(k_m,n) = \bar{\rho}(k_m),$$

where $\bar{\rho}(k)$ is increasing in k. For example, let $\rho(k,n) = k/\alpha_n$. Then $\text{plim}_{n\to\infty}\alpha_n(c_n(m) - c_n(1)) = \bar{\rho}(k_m) - \bar{\rho}(k_1) > 0$, so that

$$\lim_{n\to\infty}P[c_n(m) > c_n(1)] = 1 \text{ for } m = 2,...,K.$$

Consequently, choosing the model according to

$$\hat{m} = \underset{m=1,..,M}{\text{argmin}}\, c_n(m)$$

will asymptotically yield the most parsimonious correct model, i.e., if model 1 is this model then $\lim_{n\to\infty}P[\hat{m} = 1] = 1$.