

Lecture 10: Model Identification (continued)

Bus 41910, Time Series Analysis, Mr. R. Tsay

We now consider the problem of model selection via information criteria. There are several information criteria available in the literature. They are in the form

$$\text{crit}(m) = -2 \ln(\text{maximized likelihood}) + f(n, m)$$

where m denotes a model, n is the sample size, and $f(n, m)$ is a function of n and the number of independent parameters in the model m . Roughly speaking, the first term on the right hand side is a measure of fidelity of the model to the data (or goodness of fit) and the second term is a “penalty function” which penalizes higher dimensional models. Given a set of candidate models, the selection is typically made by choosing the model that minimizes the adopted criterion function among all candidate models.

Some of the most commonly used criterion functions for selecting ARMA(p, q) models are

- AIC: Akaike’s information criterion (Akaike, 1973)

$$\text{AIC}(p, q) = n \ln(\hat{\sigma}_a^2) + 2(p + q)$$

where $\hat{\sigma}_a^2$ is the MLE of the variance of the innovational noises. Note that for an ARMA(p, q) model, the number of independent parameters is $p + q + 2$. However, since 2 is a constant for all models, it is omitted from the above criterion function.

- BIC: Schwarz’s information criterion (Schwarz, 1978, Ann. Statist.)

$$\text{BIC}(p, q) = n \ln(\hat{\sigma}_a^2) + (p + q) \ln(n).$$

- HQ: Hannan and Quinn (1979, JRSSB)

$$\text{HQ}(p, q) = n \ln(\hat{\sigma}_a^2) + c(p + q) \ln[\ln(n)], \quad c > 2,$$

For AR(p) models, there are other criteria available:

- Akaike’s final prediction error (FPE):

$$\text{FPE}(p) = \frac{n + p}{n - p} \hat{\sigma}_p^2$$

where $\hat{\sigma}_p^2$ is the MLE of residual variance when an AR(p) model is fitted to the data.

- Akaike’s Bayesian information criterion (Bic):

$$\text{Bic}(p) = n \ln(\hat{\sigma}_z^2) - (n - p) \ln(1 - p/n) + p \ln(n) + p \ln[p^{-1}(\hat{\sigma}_z^2/\hat{\sigma}_p^2 - 1)]$$

where $\hat{\sigma}_z^2$ is the sample variance of observations. This approach is very close to the BIC of Schwarz (1978). In fact, we have

$$\text{Bic}(p) \approx \text{BIC}(p) + O(p)$$

where $O(p)$ denotes a term which is functionally independent of n .

- Parzen's CAT:

$$\text{CAT}(p) = \begin{cases} -(1 + (1/n)) & \text{if } p = 0 \\ (\frac{1}{n} \sum_{j=1}^p \frac{1}{\hat{\sigma}_j^2}) - \frac{1}{\hat{\sigma}_p^2} & \text{for } p > 0 \end{cases}$$

Recently, Hurvich and Tsai (1989, 1991, BKA) consider a bias-corrected AIC for AR(p) models as

$$\text{AICC}(p) = n \ln(\hat{\sigma}_a^2) + n \frac{1 + p/n}{1 - (p + 2)/n}.$$

This criterion function is asymptotically equivalent to AIC(p). In fact, we can write

$$\text{AICC}(p) = \text{AIC}(p) + \frac{2(p + 1)(p + 2)}{n - p - 2}.$$

This result can easily be shown by rewriting AIC(p) as

$$\text{AIC}(p) = n \ln(\hat{\sigma}_a^2) + n + 2(p + 1)$$

in which n and 2 are added. Since these two numbers are constant for all models, they do not affect the model selection. Simulation study indicates that AICC outperforms AIC in the small samples.

Discussion: Among the above criteria, BIC and HQ(.) are consistent in the sense that if the set of candidate models contains the "true" model, then these two criteria select the true model with probability 1 asymptotically. All the other criteria are inconsistent. On the other hand, since there is no "true" model in practice, "consistency" might not be a relevant property in application. Shibata (1980, Ann. Statist.) shows that AIC is asymptotically efficient in the sense that it selects the model which is closest to the unknown true model asymptotically. Here the unknown true model is assumed to be of infinite dimension.

There are advantages and disadvantages in using criterion functions in model selection. For instance, one possible disadvantage is that the selection is fully based on the data and the adopted information criterion. It is conceivable that certain substantive information is important in model selection, e.g. model interpretation. The information criterion does not incorporate such information in model selection.

In what follows, I briefly sketch a derivation of AIC information criterion. Let $f(\cdot)$ and $g(\cdot)$ be two probability density functions. A measure of goodness of fit by using $g(\cdot)$ as an estimate of $f(\cdot)$ is the entropy defined by

$$B(f; g) = - \int f(z) \ln\left(\frac{f(z)}{g(z)}\right) dz.$$

It can be shown that $B(f; g) \leq 0$ and that $B(f; g) = 0$ if and only if $f(\cdot) = g(\cdot)$. Thus, a maximum $B(f; g)$ indicates g is close to f . Akaike (1973) argues that $-B(f; g)$ can be used as a discrepancy between $f(\cdot)$ and $g(\cdot)$. Since

$$-B(f; g) = \int f(z) \ln\left(\frac{f(z)}{g(z)}\right) dz = \int \ln(f(z)) f(z) dz - \int \ln(g(z)) f(z) dz$$

$$= \text{constant} - E_f[\ln(g(z))],$$

where E_f denotes the expectation with respect to $f(\cdot)$, we define the discrepancy between $f(\cdot)$ and $g(\cdot)$ as

$$d(f; g) = E_f[-\ln(g(z))].$$

The objective then is to choose g which minimizes this discrepancy measure.

Suppose that \mathbf{x} is a set of n data points and the statistical analysis of \mathbf{x} is to predict y whose distribution is identical to that of the elements of \mathbf{x} . Such a prediction is made by using the predictive distribution of y given \mathbf{x} . Denote the true distribution of y by $f(y)$ and the predictive density of y given \mathbf{x} by $g(y|\mathbf{x})$. Then, the discrepancy is

$$d(f; g) = E_f[-\ln(g(y|\mathbf{x}))] = E_y[-\ln(g(y|\mathbf{x}))],$$

where we change the index f to y as $f(\cdot)$ is the true density function of y . This discrepancy, of course, depends on the data realization \mathbf{x} . Therefore, the expected discrepancy is

$$D(f; g) = E_x[E_y(-\ln g(y|\mathbf{x}))]$$

where E_x denotes the expectation over the joint distribution of \mathbf{x} . The question then is how to estimate this expected discrepancy.

Here $f(\cdot)$ is the true model and $g(y|\mathbf{x})$ is an entertained model. Suppose now that the entertained models $g(y|\mathbf{x})$ are indexed by the parameter θ and that the true model $f(\cdot)$ of y is within this class of candidate models, say $f(y) = g(y|\theta_0)$. Also, assume that the usual regularity conditions of MLE hold. Let $\hat{\theta}(\mathbf{x})$ be the MLE of θ given the data \mathbf{x} , i.e.

$$g(\mathbf{x}|\hat{\theta}(\mathbf{x})) = \max_{\theta} g(\mathbf{x}|\theta).$$

The following two results are well-known:

- As $n \rightarrow \infty$, the likelihood ratio statistic $2 \ln g(\mathbf{x}|\hat{\theta}(\mathbf{x})) - 2 \ln g(\mathbf{x}|\theta_0)$ is asymptotically chi-square with degrees of freedom $r = \dim(\hat{\theta}(\mathbf{x}))$.
- By Taylor expansion and asymptotic normality of MLE,

$$2 \ln g(y|\theta_0) - 2 \ln g(y|\hat{\theta}(\mathbf{x})) \approx n(\hat{\theta}(\mathbf{x}) - \theta_0)' \mathbf{I}(\hat{\theta}(\mathbf{x}) - \theta_0) \sim \chi_r^2,$$

where \mathbf{I} is the Fisher information matrix of θ evaluated at θ_0 .

Consequently, we have

$$2E_x \ln g(\mathbf{x}|\hat{\theta}(\mathbf{x})) - 2E_x \ln g(\mathbf{x}|\theta_0) = r$$

and

$$2E_x E_y \ln g(y|\theta_0) - 2E_x E_y \ln g(y|\hat{\theta}(\mathbf{x})) = r.$$

Summing over the above two equations and dividing the result by 2, we have

$$E_x \ln g(\mathbf{x}|\hat{\theta}(\mathbf{x})) - E_x E_y \ln g(y|\hat{\theta}(\mathbf{x})) = r.$$

Therefore,

$$E_x E_y [-\ln g(y|\hat{\theta}(\mathbf{x}))] = E_x [-\ln g(\mathbf{x}|\hat{\theta}(\mathbf{x}))] + r.$$

Since $E_x \ln g(\mathbf{x}|\hat{\theta}(\mathbf{x}))$ is the expectation of the logarithm of the maximized likelihood of \mathbf{x} , Akaike proposes his AIC, based on the above equation, by estimating the expected discrepancy by

$$\hat{D}(f; g) = E_x [-\ln g(\mathbf{x}|\hat{\theta}(\mathbf{x}))] + r = -\ln g(\mathbf{x}|\hat{\theta}(\mathbf{x})) + r.$$

For Gaussian time series, $-\ln g(\mathbf{x}|\hat{\theta}(\mathbf{x})) = \frac{n}{2} \ln(\hat{\sigma}_a^2) + C$, where C is a function of n and 2π . Therefore, dropping the constant C and multiplying by 2, we have

$$\text{AIC}(m) = n \ln(\hat{\sigma}_a^2) + 2r$$

where r is the dimension of $\hat{\theta}(\mathbf{x})$ and m denotes the model corresponding to the density $g(\cdot|\theta)$ entertained.