

Model selection using penalty function criteria

Laimonis Kavalieris
University of Otago
Dunedin, New Zealand

Econometrics, Time Series Analysis, and Systems Theory
Wien, June 18 – 20

Outline

- ▶ **Classes of models.**
- ▶ Penalty function model selection criteria.
 - ▶ AIC (Akaike, 1973), BIC (Schwarz, 1978), DIC, FIC, SIC, TIC and more alphabet soup.
 - ▶ MDL (Rissanen, 1979, ...)
- ▶ Analysis from the distant past – AR model under 4th moment conditions.
- ▶ Relaxation of moment conditions to $2 < \alpha < 4$ moments.
- ▶ Applications:
 - ▶ AR modelling.
 - ▶ Long memory – Fractionally integrated AR models.
 - ▶ Counting structural breaks.

Outline

- ▶ Classes of models.
- ▶ Penalty function model selection criteria.
 - ▶ AIC (Akaike, 1973), BIC (Schwarz, 1978), DIC, FIC, SIC, TIC and more alphabet soup.
 - ▶ MDL (Rissanen, 1979, ...)
- ▶ Analysis from the distant past – AR model under 4th moment conditions.
- ▶ Relaxation of moment conditions to $2 < \alpha < 4$ moments.
- ▶ Applications:
 - ▶ AR modelling.
 - ▶ Long memory – Fractionally integrated AR models.
 - ▶ Counting structural breaks.

Outline

- ▶ Classes of models.
- ▶ Penalty function model selection criteria.
 - ▶ AIC (Akaike, 1973), BIC (Schwarz, 1978), DIC, FIC, SIC, TIC and more alphabet soup.
 - ▶ MDL (Rissanen, 1979, ...)
- ▶ Analysis from the distant past – AR model under 4th moment conditions.
- ▶ Relaxation of moment conditions to $2 < \alpha < 4$ moments.
- ▶ Applications:
 - ▶ AR modelling.
 - ▶ Long memory – Fractionally integrated AR models.
 - ▶ Counting structural breaks.

Outline

- ▶ Classes of models.
- ▶ Penalty function model selection criteria.
 - ▶ AIC (Akaike, 1973), BIC (Schwarz, 1978), DIC, FIC, SIC, TIC and more alphabet soup.
 - ▶ MDL (Rissanen, 1979, ...)
- ▶ Analysis from the distant past – AR model under 4th moment conditions.
- ▶ Relaxation of moment conditions to $2 < \alpha < 4$ moments.
- ▶ Applications:
 - ▶ AR modelling.
 - ▶ Long memory – Fractionally integrated AR models.
 - ▶ Counting structural breaks.

Outline

- ▶ Classes of models.
- ▶ Penalty function model selection criteria.
 - ▶ AIC (Akaike, 1973), BIC (Schwarz, 1978), DIC, FIC, SIC, TIC and more alphabet soup.
 - ▶ MDL (Rissanen, 1979, ...)
- ▶ Analysis from the distant past – AR model under 4th moment conditions.
- ▶ Relaxation of moment conditions to $2 < \alpha < 4$ moments.
- ▶ Applications:
 - ▶ AR modelling.
 - ▶ Long memory – Fractionally integrated AR models.
 - ▶ Counting structural breaks.

Model set

Models live in a union of parametric classes \mathcal{M}_h

$$\mathcal{M} = \bigcup_{h \in H} \mathcal{M}_h$$

where $\dim(\mathcal{M}_h) = d_h$ and often $d_h < d_{h+1}$.

The simplest example is AR modelling of a stationary time series.
Then

\mathcal{M}_h consists of causal AR(h) models of the form $\sum_{j=0}^h \phi_{h,j} y_{t-j} = \epsilon_t$ where $\phi_{h,0} = 1$ and, to avoid ambiguities, $\phi_{h,h} \neq 0$.

We do not assume that the true model is in any of the \mathcal{M}_h , however it may be in the closure of \mathcal{M} . For (regular) stationary series the Wold decomposition gives a causal AR(∞) representation.

Model selection criteria

Given observations $\mathbf{y}_n = y_1, \dots, y_n$ take an estimation criterion

$$\gamma(\mathbf{y}_n; \theta)$$

For each h we can find a $\hat{\theta}_h$ minimizing $\gamma(\mathbf{y}_n; \theta), \theta \in \mathcal{M}_h$. From all of the $\hat{\theta}_h$ we would like to select a 'best' model. Define a loss function $\ell(\theta_1, \theta_2)$. A best model should minimise the expected loss.

Writing θ° as the true model, which need not be in \mathcal{M} , select h_n to minimize

$$E_{\mathbf{y}}\{\ell(\theta^\circ, \hat{\theta}_{h_n})\}$$

Of course θ° is unknown, but any data-based model selection procedure should mimic this choice (this choice is occasionally called an 'oracle').

Maximum likelihood

Minimize the negative log likelihood $\gamma(\mathbf{y}_n; \theta) = -\log f(\mathbf{y}_n; \theta)$ over \mathcal{M}_h to estimate $\hat{\theta}_h$. Then an appropriate loss function is Kullback-Liebler divergence

$$D(\theta^\circ || \hat{\theta}) = \int f(z; \theta^\circ) \log[f(z; \theta^\circ)/f(z; \hat{\theta})] dz$$

and the best model should minimize

$$E_{\mathbf{y}}\{D(\theta^\circ || \hat{\theta}_h)\} = \int f(z; \theta^\circ) \log f(z; \theta^\circ) - E\left\{ \int f(z; \theta^\circ) \log f(z; \hat{\theta}_h) \right\}$$

The first term of the RHS is independent of h so select h to minimize the second term.

Akaike's argument suggests that $-\log f(\mathbf{y}_n; \hat{\theta}_h)$ is a biased estimate of the expected K-L divergence, and the bias is $2 \times$ [number of parameters].

Least squares prediction

For 1-step prediction, minimize MSE

$$\sum_t [y_t + \sum_{j=1}^h a_j y_{t-j}]^2$$

obtaining coefficients $\hat{\phi}(h) = [\hat{\phi}_{h,1}, \dots, \hat{\phi}_{h,h}]$ and prediction errors denoted by $\hat{\epsilon}_{h,t}$. Denote the 'true' AR(∞) model by

$$y_t + \sum_{j=1}^{\infty} \phi_j y_{t-j} = \epsilon_t$$

$E\{\epsilon_t^2\} = \sigma^2$ and $\phi = [\phi_1, \phi_2, \dots]$. The loss function is

$$E\{(\hat{\epsilon}_{h,t} - \epsilon_t)^2\} = E\left\{\left(\sum (\hat{\phi}_{h,j} - \phi_j) y_{t-j}\right)^2\right\}$$

and the preferred model will minimize risk.

Define $\epsilon_{h,t} = y_t + \sum \phi_{h,j} y_{t-j}$ as the (unobservable) prediction error obtained by minimizing $E\{[y_t - \sum_{j=1}^h a_j y_{t-j}]^2\}$.

Least square prediction ctd. . .

Evidently

$$\hat{\epsilon}_{h,t} - \epsilon_t = (\hat{\epsilon}_{h,t} - \epsilon_{h,t}) + (\epsilon_{h,t} - \epsilon_t)$$

Then the expected loss becomes the sum of two squared terms

$$E\{E\{(\hat{\epsilon}_{h,t} - \epsilon_t)^2\}\} = E\{E\{(\hat{\epsilon}_{h,t} - \epsilon_{h,t})^2\}\} + \|\phi(h) - \phi\|_F^2$$

The first term of the loss function is σ^2/n (it is the expectation of a χ_h^2 distribution) – this term accounts for the variance in estimating the h parameters in the autoregression model. The second term is the bias due to using the wrong model – decreases with increasing h . Let h_n minimise the expected loss.

As there is no true model in \mathcal{M} we cannot talk about consistency, so a good data-based model selection procedure should select model close to $\hat{\theta}_{h_n}$.

Penalty function criteria

Penalize the estimation criterion according to the number of parameters. So, select the model that minimizes

$$\gamma(\mathbf{y}_n; \hat{\theta}_h) + p(h, n)$$

Depending on the kind of model being considered, every parameter need not have the same penalty. [When estimating a periodic signal, amplitudes are estimated with variance $O(n^{-1})$ while frequencies are estimated with much higher precision, variance $O(n^{-3})$. This requires different penalties, MDL for example suggests a penalty of $\log n/n$ for each amplitude, and $3 \log n/n$ for each frequency, (Kavalieris and Hannan, 1993). Similar considerations apply when estimating the number of breakpoints, though that depends on the choice of asymptotics.]

Technical results from the 80's

Basic results used for the analysis of are the uniform convergence results for autocovariances: Denote sample and population covariance by $\hat{\gamma}(h) = n^{-1} \sum_{j=h+1}^n y_t y_{t-k}$ and $\gamma(h) = E\{y_t y_{t-h}\}$ respectively. Then

$$\max_{0 \leq h \leq H_n} |\hat{\gamma}(h) - \gamma(h)| = O \left\{ \frac{\log n}{n} \right\}^{1/2}$$

almost surely. Conditions: $y_t = \sum \phi_j \epsilon_{t-j}$ where ϵ_t are stationary martingale differences with a finite 4th moment.

These results are due to Ted Hannan + coworkers; a synthesis of those results appears in Hannan and Deistler (1988).

Technical results ctd ...

For AR modelling, a penalty function criterion is

$$\log \hat{\sigma}_h^2 + h \log n/n$$

where $\hat{\sigma}_n^2 = n^{-1} \sum \hat{\epsilon}_{h,t}^2$ is the mean square prediction error from an AR(h) model.

The prediction variance is partitioned into three bits:

$$\begin{aligned} \hat{\sigma}_h^2 = \frac{1}{n} \sum \hat{\epsilon}_{h,t}^2 &= \frac{1}{n} \sum \epsilon_t^2 \\ &+ [\text{Variance due to } h \text{ parameters}] \\ &+ [\text{Bias due to wrong model}] \end{aligned}$$

Penalty term must wipe out enough variation in the 2nd and 3rd terms so that they look like the expected loss function for least squares estimation.

Results are quite complete when an AR model is used to estimate a time series with geometrically decaying ACF (the ARMA case).

Moments of order $2 < \alpha < 4$

Can we relax the moment conditions? Yes, but at a cost ...

Kavalieris (2008) shows that if $E\{|\epsilon_t|^\alpha\} < \infty$ and the ϵ_t are independent, then

$$\max_{0 \leq h \leq H_n} \left| \frac{1}{n} \sum \epsilon_t \epsilon_{t-h} \right| = O \left\{ \frac{\log n}{n} \right\}^{1/2}$$

almost surely, but this is not enough to establish a similar rate of convergence for auto correlations. Under the independence assumption, we can show that

$$\max_{0 \leq h \leq H_n} |\hat{\gamma}(h) - \gamma(h)| = O_p \left\{ \frac{\log n}{n} \right\}^{1/2}$$

in probability (and the rate given here is the best possible when $H_n \sim n^a$ for any $0 < a < 1$). Then most of the Hannan and Kavalieris (1986) results can be recovered.

Long memory

The $O(\log n/n)^{1/2}$ convergence rate was not established for long memory – it requires

$$\sum j^{1/2} |\psi_j| < \infty$$

where $y_t = \sum \psi_j \epsilon_{t-j}$. If the ϵ_t are iid and normally distributed, then we can recover the $O(\log n/n)^{1/2}$ convergence rate, and a little more. [Normality is not really needed, only many moments!]

Define $v_{k,t} = \sum \psi_{k,j} \epsilon_{t-j}$, $k \leq n^a$ for any $a > 0$ where the $\psi_{k,j}$ are square summable. Then

$$\max_{k, \ell \leq n^a} \frac{\sum [v_{k,t} v_{\ell,t} - E\{v_{k,t} v_{\ell,t}\}]}{\sqrt{\text{Var}\{v_{k,t}\} \text{Var}\{v_{\ell,t}\}}} = O(n \log n)^{1/2}$$

Here it does not matter that $\min_k \text{Var}\{v_{k,t}\} = 0$.

Fractional AR models

Model:

$$y_t = (1 - z)^{-d} \phi(z) \epsilon_t$$

where $\phi(z) = 1 + \sum_1^\infty \phi_j z^j$. This is the fractional ARIMA($\infty, d, 0$) model where the AR component is meant to capture the short range correlation structure. We want to estimate the model as an ARIMA($h, d, 0$) model

$$y_t = (1 - z)^{-d} \phi_h(z) \epsilon_t$$

where now $\phi_h(z)$ contains only h lags.

In this case estimating h by minimizing a BIC penalty criterion does lead to consistent estimates of the spectrum (and a consistent estimate of d).

References

- ▶ Hannan, E.J. and Kavalieris, L. (1986) Regression-autoregression models, *J. Time Series Analysis*, **7**, 27-49.
- ▶ Hannan, E.J. and Deistler, M. (1998) *The statistical theory of linear systems* Wiley, New York.
- ▶ Kavalieris, L. and Hannan, E.J (1994) Determining the number of terms in a trigonometrical series. *J.T.S.A.* **15**, 613 – 626.
- ▶ Kavalieris, L. (2008) Uniform convergence of autocovariances. *Statistics and Probability Letters* **78**, 830-838.