Lecture 4: Model Selection "Roundup" and Time Series Examples

Francis J. DiTraglia

March 19, 2014

1 More on Consistency vs Efficiency

Briefly explain catching up by switching sooner if there's time. Try to see what's going on with the weak consistency assumption from Sin and White as regards a penalty of zero.

2 Overview of What We know so Far and time series examples

So far we've looked at some completely generic examples (AIC, TIC, cross-validation) and some regression examples (Mallows, AIC corrected). The generic examples can immediately be applied to an ML problem, including time series setting: if you have an ML routine, you already get everything you need for these criteria as a side effect. Could be Kalman filter or conditional likelihood.

We derived the other examples (Mallows, AIC corrected) for a regression problem, but it's easy to adapt these to AR and VAR models. As long as we're willing to use conditional ML (drop some observations) these already *are* regression problems.

We'll take a look at AR and VAR models using conditional likelihood so we can write out explicit formulas for the criteria. We won't do TIC since we can't really unpack the penalty term. We'll treat cross-validation in its own section.

The consistent criteria are ... and the efficient criteria are ...

We won't go through all of the specifics of the derivations for mallows and AICc since they're almost identical to the regression derivation. Some more details can be found in McQuarrie and Tsai (1998).

2.1 Autoregressive Models

2.2 VAR Models

Pretty much the same as AR but multivariate.

2.3 Vector Autoregression Models

Write without an intercept for simplicity (just demean everything)

$$\begin{aligned} \underset{(q\times 1)}{\mathbf{y}_t} &= \underset{(q\times q)}{\Phi_1}\, \mathbf{y}_{t-1} + \ldots + \Phi_p \mathbf{y}_{t-p} + \epsilon_t \\ \boldsymbol{\epsilon}_t &\overset{iid}{\sim} N_q(\mathbf{0}, \Sigma) \end{aligned}$$

Conditional least squares estimation, sample size, etc.

$$
\begin{aligned}
FPE &= \left|\widehat{\Sigma}_p\right| \left(\frac{T+qp}{T-qp}\right)^q \\
AIC &= \log\left|\widehat{\Sigma}_p\right| + \frac{2pq^2+q(q+1)}{T} \\
AIC_c &= \log\left|\widehat{\Sigma}_p\right| + \frac{(T+qp)q}{T-qp-q-1} \\
BIC &= \log\left|\widehat{\Sigma}_p\right| + \frac{\log(T)pq^2}{T} \\
HQ &= \log\left|\widehat{\Sigma}_p\right| + \frac{2\log\log(T)pq^2}{T}
\end{aligned}
$$

Problems with VAR model selection

1. If we fit p lags, we lose p observations under the conditional least squares estimation procedure.
2. Adding a lag introduces q^2 additional parameters.

Corrected AIC for State Space Models Problem with VARs and state space more generally is that we can easily have sample size small relative to number of parameters. In this case AIC-type criteria don't work well. Suggestions for simulation-based selection.

Cavanaugh & Shumway (1997)

3 More on Cross-Validation

How to extend it to time series. Varieties other than leave-one-out. Efficiency versus consistency. Racine (2000) and Burman, Chow & Nolan (1994).

3.1 How to handle dependent observations

AR example.

Cross-Validation for AR The way we described it above, CV depended in independence. How can we adapt it for AR models? Roughly speaking, the idea is to use the fact that dependence dies out over time and treat observations that are "far enough apart" as *approximately* independent. Specifically, we choose an integer value h and assume that y_t and y_s can be treated as independent as long as $|s-t| > h$. This idea is called "h-block cross-validation" and was introduced by Burman, Chow & Nolan (1994). As in the iid version of leave-one-out cross-validation, we still evaluate a loss function by predicting *one* witheld observation at a time using a model estimated without it. The difference is that we also omit the h neighboring observations *on each side* when fitting the model. For example, if we choose to evaluate squared-error loss, the criterion is

$$CV_h(1) = \frac{1}{T-p} \sum_{t=p+1}^{T} \left(y_t - \hat{y}^h_{(t)}\right)^2$$

where

$$\hat{y}^h_{(t)} = \hat{\phi}^h_{1(t)} y_{t-1} + \ldots + \hat{\phi}^h_{1(t)} y_{t-p}$$

and $\hat{\phi}^h_{j(t)}$ denotes the jth parameter estimate from the conditional least-squares estimator with observations $y_{t-h}, \ldots, y_{t+h}$ removed. We still have the question of what h to choose. Here there is a trade-off between making the assumption of independence more plausible and leaving enough observations to get precise model estimates. Intriguingly, the simulation evidence presented in McQuarrie and Tsai (1998) suggests that setting $h = 0$, which yields plain-vanilla leave-one-out CV, works well even in settings with dependence.

The idea of h-block cross-validation can also be adapted to versions of cross-validation other than leave-one-out. For details, see Racine (1997, 2000).

Cross-Validation for VARs In principle we could use the same h-block idea here as we did for for the AR example above. However, given the large number of parameters we need to estimate, the sample sizes witholding $2h+1$ observations at a time may be too small for this to work well.

4 Two Additional Criteria

One efficient another consistent.

4.1 Final Prediction Error

4.2 Hannan-Quinn