Lecture 3: Asymptotic Properties

Francis J. DiTraglia

March 9, 2014

1 Time Series Examples

We won't go through all of the specifics here since they're almost identical to the material from above. Some more details can be found in McQuarrie and Tsai (1998). The AR and VAR models are straightforward since, in the conditional formulation, they're just univariate and multivariate regression, respectively.

1.1 Autoregressive Models

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).

1.2 Vector Autoregression Models

Write without an intercept for simplicity (just demean everything)

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

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.

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.

1.3 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)

2 Consistency vs. Efficiency

2.1 Introduction

Up until now we've made proceeded by setting forth desiderata for model selection, e.g. minimize the KL divergence or predictive mean-squared error, and then making enough assumptions until we could derive a criterion. And although the details of the derivations were all different, in each of the examples we've considered to far, the result amounted to adding a penalty to the maximized log-likelihood to account for model complexity, for example:

$$\begin{aligned} TIC &= 2\ell_T(\widehat{\theta}) - \text{trace}\left\{\widehat{J}^{-1}\widehat{K}\right\} \\ AIC &= 2\ell_T(\widehat{\theta}) - 2\ \text{length}(\theta) \\ BIC &= 2\ell_T(\widehat{\theta}) - \log(T)\ \text{length}(\theta) \end{aligned}$$

We're now going to take a completely different perspective. Instead of asking what assumptions we need to derive a particular criterion, we'll ask "given the penalty term that this criterion applies to the log-likelihood, how will it perform in large samples?" We'll concern ourselves in particular with two properties: **consistency** and **efficiency**.

Consistency Suppose that we have a set of candidate models, one of which is actually the true DGP. It seems clear that in this setting we'd like our model selection procedure to correctly identify the true DGP as the sample size grows. This is the idea behind consistency. We say that a model selection criterion is **consistent** if it selects the true DGP with probability approaching one as $T \to \infty$.

Efficiency It's somewhat rare that the goal of model selection is to determine which model is the "truth" or even which model is the KL minimizer. More commonly we estimate a model for *some specific purpose*: perhaps we want to estimate a particular parameter or make a good forecast. From this

perspective it is natural to look for a model selection criterion that with good risk properties. Intuitively, we'l like the criterion to perform "almost as well" as the risk-optimal model in our candidate set. This property, which we'll make more precise below, is called **efficiency**.

You may be thinking "consistency and efficiency both sound like great properties so let's find a criterion that satisfies them both!" Unfortunately, this turns out to be impossible: if a model selection criterion is consistent it cannot be efficient, and vice-versa.

Weak Consistency But what if the true DGP is not among the candidate models? This seems like a much more realistic assumption. If we are willing to assume that there is a unique candidate model with minimum KL divergence from the truth then it makes sense to ask that our model selection criterion identify *this model* as the sample size grows. We say that a model selection criterion is **weakly consistent** if it selects the KL minimizing candidate model with probability approaching one as $T \to \infty$.