Lecture 3: Asymptotic Properties

Francis J. DiTraglia

March 17, 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:

$$\begin{aligned} AIC &= 2\ell_T(\widehat{\theta}) - 2 \text{ length}(\theta) \\ BIC &= 2\ell_T(\widehat{\theta}) - \log(T) \text{ length}(\theta) \end{aligned}$$

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.

2.2 Conditions

Let g be the true, unknown data density and consider a collection of models M_k indexed by $k = 1, 2, \ldots, K$ where θ_k is the parameter vector under model M_k and $\widehat{\theta_k}$ is the corresponding maximum likelihood estimator. Let $f_{k,t}(y_t|\theta_k)$ be the density of observation t under model k. For simplicity, suppose that we can express the likelihood of model k as $\sum_{t=1}^{T} \log f_{k,t}(Y_t|\theta_k)$. This isn't actually necessary: if you want to see a more general way of writing things, consult Sin and White (1996). We do *not* assume that the data are independent. Suppose we're interested in choosing a model to mininimze the KL divergence from g to f_k.

General Form of Information Criteria

$$IC(M_k) = 2\sum_{t=1}^{T} \log f_{k,t}(Y_t|\widehat{\theta_k}) - c_{T,k}$$

where $c_{T,k}$ is the penalty term for M_k. We'll now ask how different choices of $c_{T,k}$ give rise to criteria that behave in different ways.

2.3 Weak Consistency

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 $T \to \infty$.

Sufficient Conditions for Weak Consistency Suppose that exactly one of the candidates minimizes the KL distance: call it M_{k_0}. To state this precisely, suppose that

$$\liminf_{T\to\infty} \left(\min_{k \neq k_0} \frac{1}{T} \sum_{t=1}^{T} \{KL(g; f_{k,t}) - KL(g; f_{k_0,t})\} \right) > 0$$

Then, if $c_{T,k} > 0$ and $c_{T,k} = o_p(T)$, $IC(M_k)$ is *weakly consistent*: it selects M_{k_0} with probability approaching one in the limit. Weak consistency continues to hold if the penalty term $c_{T,k}$ equals zero for one of the models, so long as it is strictly positive for all of the others.

Both AIC and BIC are Weakly Consistent We have

$$\begin{aligned} \text{BIC Penalty:} \quad & c_{T,k} = \log(T) \times \text{length}(\theta_k) \\ \text{AIC Penalty:} \quad & c_{T,k} = 2 \times \text{length}(\theta_k) \end{aligned}$$

and both of these penalties satisfy the condition $T^{-1} c_{T,k} \xrightarrow{p} 0$.

2.4 Consistency

But what if *two or more* models minimize the KL-divergence? We very often use information criteria to select among *nested models* to decide, for example, whether to restrict certain elements of θ to be equal to zero. Suppose we want to choose the number of lags to include in an AR model. The usual way to do this is to specify a maximum lag-length, say 3 periods, and then evaulate each of the AR models up to this order: AR(1), AR(2), and AR(3). But in this

example is is entirely possible that the KL minimizer will *fail* to be unique. The AR(2) model is just a special case of the AR(3) with one coefficient set equal to zero. Similarly, the AR(1) model is just a special case of the AR(2). Stated mode generally, if an AR(k) model with all coefficients different from zero is the KL minimizer, then an AR(k+1) model also minimizes the KL divergence, as does an AR(k+2) and an AR(k+3) by setting certain coefficients to zero. In situations like this, where there is a tie in the KL divergence, it makes sense to choose the most "parsimonious" specification, in other words the one with the fewest parameters. This idea is often called **consistency**.

Sufficient Conditions for Consistency Suppose that, among our set of candidate models there is a tie in the KL divergence. Let $\mathcal{J}$ be the set of all models that attain the minimum KL divergence. Among these, let $\mathcal{J}_0$ denote the subset with the minimum number of parameters. *Either* of the following two conditions is sufficient for consistency. In other words, both (a) and (b) imply that we will select a model from $\mathcal{J}_0$ with probability approaching one in the limit:

$$\underset{T\to\infty}{P}\left\{\min_{\ell\in\mathcal{J}\backslash\mathcal{J}_0}\left[IC(M_{j_0})-IC(M_\ell)\right]>0\right\}=1$$

Here are the alternative sets of conditions:

(a) The following two conditions are sufficient for consistency:

(i) For all $k\neq\ell\in\mathcal{J}$

$$\limsup_{T\to\infty}\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left\{KL(g;f_{k,t})-KL(g;f_{\ell,t})\right\}<\infty$$

(ii) For all $j_0\in\mathcal{J}_0$ and $\ell\in(\mathcal{J}\backslash\mathcal{J}_0)$

$$P\left\{\left(c_{T,\ell}-c_{T,j_0}\right)/\sqrt{T}\to\infty\right\}=1$$

(b) The following two conditions are *also* sufficient for consistency:

(i) For all $k \neq \ell \in \mathcal{J}$

$$\sum_{t=1}^{T} \left[\log f_{k,t}(Y_t|\theta_k^*) - \log f_{\ell,t}(Y_t|\theta_\ell^*)\right] = O_p(1)$$

where θ_k^* and θ_ℓ^* are the respective KL minimizing parameter values.

(ii) For all $j_0 \in \mathcal{J}_0$ and $\ell \in (\mathcal{J} \backslash \mathcal{J}_0)$

$$P\left(c_{T,\ell} - c_{T,j_0} \to \infty\right) = 1$$

Note that each of these alternative sets of conditions has *two parts*: the first is a regularity condition that restricts the asymptotic behavior of the models in $\mathcal{J}$ while the second is a condition on the penalty term $c_{T,k}$. We immediately see that the penalty terms for the AIC and TIC *cannot* satisfy (a)(ii) or (b)(ii) since $(c_{T,\ell} - c_{T,j_0})$ *does not depend on sample size.* While this does not consitute a proof, it does turn out that neither is consistent: even in the limit AIC and TIC have a non-zero probability of "overfitting," i.e. selection a model that is in $\mathcal{J} \backslash \mathcal{J}_0$. In constrast, under (b)(i) the BIC *is consistent* since

$$c_{T,\ell} - c_{T,j_0} = \log(T)\left\{\text{length}(\theta_\ell) - \text{length}(\theta_{j_0})\right\}$$

The term in braces is *positive* since $\ell \in \mathcal{J} \backslash \mathcal{J}_0$, i.e. ℓ is not as parsimonious as j_0, and $\log(T) \to \infty$. This means that in the limit, BIC will *always* select a model in $\mathcal{J}_0$.

What about selecting the true DGP? The way we will just defined consistency did *not* in fact require that the true DGP is among the models under consideration. If the true DGP *is* among the models in our set, however, the preceding result gives conditions under which we are guaranteed to select it in the limit. Why is this the case? First of all, the true DGP minimizes the KL and the minimized value is zero. (See the notes for Lecture 1.) The only way that *another* model could also minimize the KL divergence in this case is

if it has "superfluous" parameters. For example, suppose the true DGP is an AR(1) but we also consider an AR(2). Hence, the true DGP is necessarily the most parsimonious model among those that minimize the KL divergence.

2.5 The Simplest Possible Example

Let $Y_1, \ldots, Y_T \overset{\text{iid}}{\sim} N(\mu, 1)$ and consider two models: M_0 assumes that $\mu = 0$ while M_1 doesn't make any assumption about the value of μ. Now suppose we want to use an information criterion to choose between M_0 and M_1. We'll consider penalty terms of the form $c_{T,k} = d_T \times \text{length}(\theta_k)$ which includes both the AIC and BIC as special cases. Since M_0 has *zero* parameters while M_1 has one parameter, our information criteria are as follows:

$$\begin{aligned} IC_0 &= 2 \max_{\mu} \{\ell_T(\mu) \colon M_0\} \\ IC_1 &= 2 \max_{\mu} \{\ell_T(\mu) \colon M_1\} - d_T \end{aligned}$$

$$\begin{aligned} \ell_T(\mu) &= \sum_{t=1}^{T} \log \left(\frac{1}{2\pi} \exp \left\{ -\frac{1}{2} (Y_t - \mu)^2 \right\} \right) \\ &\vdots \quad \boxed{\text{fill in later}} \\ &= -\frac{T}{2} \left\{ \widehat{\sigma}^2 + \log(2\pi) \right\} - \frac{T}{2} \left(\bar{Y} - \mu \right)^2 \\ &= C - \frac{T}{2} \left(\bar{Y} - \mu \right)^2 \end{aligned}$$

Hence, substituting 0 for μ under M_0 and the MLE $\bar{Y}$ for μ under M_1, we have

$$\begin{aligned} IC_0 &= 2 \max_{\mu} \{\ell_T(\mu) \colon M_0\} = 2C - T\bar{Y}^2 \\ IC_1 &= 2 \max_{\mu} \{\ell_T(\mu) \colon M_1\} - d_T = 2C - d_T \end{aligned}$$

Therefore,

$$IC_1 - IC_0 = T\bar{Y}^2 - d_T$$

and we choose M_1 if this quantity is positive, in other words if

$$\begin{aligned} T\bar{Y}^2 &\geq d_T \\ \left|\sqrt{T}\bar{Y}\right| &\geq \sqrt{D_T} \\ \sqrt{T}\left|\bar{Y}\right| &\geq \sqrt{d_T} \end{aligned}$$

Thus, our selected model is

$$\widehat{M} = \begin{cases} M_1, & \sqrt{T}\left|\bar{Y}\right| \geq \sqrt{d_T} \\ M_0, & \sqrt{T}\left|\bar{Y}\right| < \sqrt{d_T} \end{cases}$$

and our *post-selection estimator* is

$$\widehat{\mu} = \begin{cases} \bar{Y}, & \sqrt{T}\left|\bar{Y}\right| \geq \sqrt{d_T} \\ 0, & \sqrt{T}\left|\bar{Y}\right| < \sqrt{d_T} \end{cases}$$

For the AIC we have $d_T = 2$ while for the BIC we have $d_T = \log(T)$. Now let's examine the asymptotics as $T \to \infty$.

Case I: $\mu \neq 0$ In this case, M_1 is the true DGP and the unique KL-minimizer. Since it's the true DGP, the KL divergence for M_1 is exactly zero. (See Lecture 1.) We can calculate the KL divergence for M_0 using similar steps to those we employed to derive the AIC_c in Lecture 2.

fill in details later

To summarize, we have

$$\begin{aligned} KL(M_1) &= 0 \\ KL(M_0) &= \frac{\mu^2}{2} \end{aligned}$$

Now let's check our sufficient conditions for weak consistency. First, we have

$$\begin{aligned}\liminf_{T\to\infty}\left(\min_{k\neq k_0}\frac{1}{T}\sum_{t=1}^{T}\{KL(g;f_{k,t})-KL(g;f_{k_0,t})\}\right) &= \liminf_{n\to\infty}\frac{1}{T}\sum_{t=1}^{T}\left(\frac{\mu^2}{2}-0\right)\\ &= \liminf_{T\to\infty}\left(\frac{\mu^2}{2}\right)>0\end{aligned}$$

as required. Now, the condition on the penalty term is $c_{T,k}=o_p(T)$, in other words $c_{T,k}/T \xrightarrow{p} 0$ both the AIC and BIC penalties satisfy this condition. Hence, if M_1 is the true model, both the AIC and BIC will select it with probability approaching 1 as $T\to\infty$.

Case II: $\mu=0$ In this case, both M_1 and M_0 are true and *both* minimize the KL divergence. The most parsimonious model, however, is M_0. Hence, using our notion of consistency (*not* weak consistency), we'd like our criteria to select M_0. We'll use the second set of sufficient conditions for consistency. In this example, it's easy to verify (b)(i). Since a $N(0,1)$ model is nested inside a $N(\mu,1)$ model, if the true distribution is $N(0,1)$ then the likelihood ratio statistic is asymptotically $\chi^2(1)$, hence the log-likelihood ratio is $O_p(1)$ as required. We know from above that the AIC penalty does *not* satisfy (b)(ii) but the BIC penalty *does*. Hence the BIC will select M_0 with probability approaching one in the limit.

Finite Sample Selection Probabilities Since this is such a simple example, we can do better than appeal to asymptotics: we can calculate the exact finite-sample behavior of the selection criteria. The AIC penalty is $2\times\text{length}(\theta)$ which corresponds to $d_T=2$. Hence, the AIC-selected model is

$$\widehat{M}_{AIC}=\begin{cases}M_1, & |\sqrt{T}\bar{Y}|\geq\sqrt{2}\\ M_0, & |\sqrt{T}\bar{Y}|<\sqrt{2}\end{cases}$$

Hence,

$$\begin{aligned} P\left(\widehat{M}_{AIC} = M_1\right) &= P\left(\left|\sqrt{T}\bar{Y}\right| \geq \sqrt{2}\right) \\ &= P\left(\left|\sqrt{T}\mu + Z\right| \geq \sqrt{2}\right) \\ &= P\left(\sqrt{T}\mu + Z \leq -\sqrt{2}\right) + \left[1 - P\left(\sqrt{T}\mu + Z \leq \sqrt{2}\right)\right] \\ &= \Phi\left(-\sqrt{2} - \sqrt{T}\mu\right) + \left[1 - \Phi\left(\sqrt{2} - \sqrt{T}\mu\right)\right] \end{aligned}$$

where $Z \sim N(0,1)$ using the fact that $\bar{Y} \sim N(\mu, 1/T)$ since $Var(Y_t) = 1$.

Now, the BIC penalty is $\log(T) \times \text{length}(\theta)$ which corresponds to $d_T = \log(T)$. Hence, the BIC-selected model is

$$\widehat{M}_{BIC} = \begin{cases} M_1, & |\sqrt{T}\bar{Y}| \geq \sqrt{\log(T)} \\ M_0, & |\sqrt{T}\bar{Y}| < \sqrt{\log(T)} \end{cases}$$

Using the exact same steps as for the AIC except with $\sqrt{\log(T)}$ in the place of $\sqrt{2}$, we have

$$\begin{aligned} P\left(\widehat{M}_{BIC} = M_1\right) &= P\left(\left|\sqrt{T}\bar{Y}\right| \geq \sqrt{\log(T)}\right) \\ &= \Phi\left(-\sqrt{\log(T)} - \sqrt{T}\mu\right) + \left[1 - \Phi\left(\sqrt{\log(T)} - \sqrt{T}\mu\right)\right] \end{aligned}$$

Replicate Figure 4.1 from the book in R. Perhaps make a Shiny App if there's time.

What is the probability of overfitting? The case where $\mu = 0$

For a generic information criterion of the form we're considering here:

$$P\left(\widehat{M} = M_1\right) = P\left(|\sqrt{T}\bar{Y}| \geq \sqrt{d_T}\right)$$