Lecture 3: Asymptotic Properties

Francis J. DiTraglia

March 18, 2014

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.

In this lecture we'll consider the following basic setup. 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 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 $T^{-1}c_{T,k} \xrightarrow{p} 0$.

3 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:

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

Here are the alternative sets of conditions:

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

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

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

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

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

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

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

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

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

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

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.

4 Efficient Model Selection

Roughly speaking, a model selection criterion is called efficient if it performs "nearly as well" as the theoretical optimum relative to some loss function. To make this more concrete, we'll look at a particular example.

Let $\{\epsilon_t\}_{-\infty}^{\infty}$ by an iid sequence of $N(0,\sigma^2)$ random variables and let $\{X_t\}$ be a stationary Gaussian Process that satisfies

$$X_t + a_1 X_{t-1} + a_2 X_{t-2} + \cdots = \epsilon_t$$

for some set of coefficients $\{a_j\}$. We attempt to approximate this stochastic process with an AR(k) model, namely

$$X_t + a_1 X_{t-1} + \cdots + a_2 X_{t-k} = \epsilon_t$$

and calculate estimates $\widehat{a}_1, \ldots, \widehat{a}_k$ using observations $X_1, \ldots, X_T$. Now, suppose our goal is to make good one-step-ahead forecasts where "good" means minimum mean-squared prediction error. To keep things simple it is typically assumed that we have a *new* realization $Y_1, \dots, Y_T$ of the *same* time series that is independent of $X_1, \dots, X_T$. This is indeed an unrealistic assumption, but it simplifies various calculations. Although it's possible to proceed without it, you'll often see it invoked. The one-step-ahead prediction is

$$\widehat{Y}_{t+1} = -\widehat{a}_1 Y_t - \widehat{a}_2 Y_{t-2} - \cdots - \widehat{a}_k Y_{t-k+1}$$

as the one-step-ahead mean-squared prediction error is

$$MSPE(k) \;\; = \;\; E\left[\left(Y_{t+1} - \widehat{Y}(k)_{t+1}\right)^2 | \, X_1, \dots, X_T\right]$$

Our ideal would be to estimate an $AR(k^*)$ model for forecasting where k^* minimizes $MSPE(k)$. Since we don't know k^* we use a model selection criterion to estimate it. Let $\widehat{k}$ be the lag-length that is selected by some model selection criterion. We say that this criterion is *asymptotically efficient* if

$$\frac{MSPE(\widehat{k})}{MSPE(k^*)} \xrightarrow{p} 1 \quad \text{as} \quad T \to \infty$$

Under appropriate assumptions, it can be shown for this example that the AIC and AIC$_c$ are asymptotically efficient while the BIC is not.

5 A Simple Example

Let $Y_1, \dots, Y_T \overset{\text{iid}}{\sim} N(\mu, 1)$ 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:

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

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

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

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

Thus, our selected model is

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

and our *post-selection estimator* is

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

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. First, we have

$$\begin{aligned} \log g(y) - \log f_\theta(y) &= -\frac{1}{2}(y-\mu)^2 + \frac{1}{2}y^2 \\ &\vdots \quad \boxed{\text{a little algebra}} \\ &= \mu\left(y - \frac{\mu}{2}\right) \end{aligned}$$

since the constant $-\log(2\pi)/2$ appears in each term and hence cancels. Thus,

$$\begin{aligned} KL(g; f_\theta) &= \int \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(y-\mu)^2\right\} \mu\left(y - \frac{\mu}{2}\right)\, dy \\ &\vdots \quad \boxed{\text{fill in later}} \\ &= \frac{\mu^2}{2} \end{aligned}$$

To summarize, we have

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

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 $c_{T,k}/T \overset{p}{\to} 0$ 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

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

Shiny App: `http://glimmer.rstudio.com/fditraglia/CH_Figure_4_1/`

What is the probability of overfitting? Suppose $\mu = 0$. In this case both models are KL-minimizers but we'd prefer M_0 since it's more parsimonious. For a generic information criterion of the form we're considering here, we calculate the "probability of overfitting" as follows

$$\begin{aligned}
P\left(\widehat{M} = M_1\right) &= P\left(|\sqrt{T}\bar{Y}| \geq \sqrt{d_T}\right) = P(|Z| \geq \sqrt{d_T})\\
&= P(Z^2 \geq d_T) = P(\chi^2_1 \geq d_T)
\end{aligned}$$

where $Z \sim N(0,1)$. For the AIC $d_T = 2$ so the probability of overfitting is $P(\chi^2_1 \geq 2) \approx 0.157$. For the BIC $d_T = \log(T)$ so the probability of overfitting is $P(\chi^2_1 \geq \log T) \rightarrow 0$ as $T \rightarrow 0$.

The Post-Selection Estimator

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

Consider MSE risk, scaling up by T since variances for well-behaved problems are $O(1/T)$. Recall from above that $\sqrt{T}\bar{Y} = \sqrt{T}\mu + Z$ where $Z \sim N(0,1)$. Thus,

$$\begin{aligned} R_T(\mu) &= TE_\mu\left[(\widehat{\mu} - \mu)\right] = E_\mu\left[\left(\sqrt{T}\widehat{\mu} - \sqrt{T}\mu\right)\right] \\ &= E\left\{\left[\left(\sqrt{T}\mu + Z\right)\mathbf{1}\left\{\sqrt{T}\mu + Z \geq \sqrt{d_T}\right\} - \sqrt{T}\mu\right]^2\right\} \\ &\vdots \quad \boxed{\text{fill in later}} \\ &= 1 - \int_a^b z^2\phi(z)\,dz + T\mu^2\left[\Phi(b) - \Phi(a)\right] \end{aligned}$$

where

$$\begin{aligned} a &= -\sqrt{d_T} - \sqrt{T}\mu \\ b &= \sqrt{d_T} - \sqrt{T}\mu \end{aligned}$$

To evaulate this risk function, we need an explicit formula for the integral that makes up the second term. This sounds like a job for integration by parts! We'll take $u = -z$ and $dv = -z\exp\{-z^2/2\}$ since

$$\frac{d}{dz}\left(\exp\left\{-\frac{z^2}{2}\right\}\right) = -z\exp\left\{-\frac{z^2}{2}\right\}$$

Thus, $v = \exp\{-z^2/2\}$, $du = -1$ and we have

$$\begin{aligned}
\int_a^b z^2 \phi(z)\, dz &= \frac{1}{\sqrt{2\pi}} \int_a^b z^2 \exp\left\{-\frac{z^2}{2}\, dz\right\} \\
&= \frac{1}{\sqrt{2\pi}} \left[-z \exp\left\{-\frac{z^2}{2}\right\}\Bigg|_a^b + \int_a^b \exp\left\{-\frac{z^2}{2}\right\}\, dz \right] \\
&= a\phi(a) - b\phi(b) + \Phi(b) - \Phi(a)
\end{aligned}$$

Putting it all together, we have

$$\begin{aligned}
R_T(\mu) &= 1 - [a\phi(a) - b\phi(b) + \Phi(b) - \Phi(a)] + T\mu^2 [\Phi(b) - \Phi(a)] \\
&= 1 + [b\phi(b) - a\phi(a)] + (T\mu^2 - 1)[\Phi(b) - \Phi(a)]
\end{aligned}$$

where

$$\begin{aligned}
a &= -\sqrt{d_T} - \sqrt{T}\mu \\
b &= \sqrt{d_T} - \sqrt{T}\mu
\end{aligned}$$

Shiny App: `http://glimmer.rstudio.com/fditraglia/CH_Figure_4_2/`

What we see from the picture is that, in exchange for low risk in a small neighborhood of $\mu = 0$, BIC has max risk that *diverges* as the sample size increases. In contrast, AIC has bounded max risk. This isn't just a problem with BIC: *any* consistent selection criterion will suffer from this defect.

Postscript The preceding is a special example of a more general phenomenon: consistency and efficiency are mutually exclusive properties. In general, consistent model selection criterion will have unbounded minimax risk. There is a huge literature on this topic but it's fairly technical. The key observation is that pointwise and uniform risk approximations give very different results. Yang (2007) gives a readable introduction. See Leeb and Pötscher (2009) for a comprehensive list of references.