Lecture 1: "AIC-type" Information Criteria

Francis J. DiTraglia

March 4, 2014

1 Introduction

Akaike's Information Criterion (AIC) summarizes the quality of a model by trading fit, measured by the maximized log likelihood, against complexity, measured by the number of estimated parameters. But where does this complexity penalty come from? In these notes we'll take a closer look at the AIC along with two related information criteria: Takeuichi's Information Criterion (TIC), and the Corrected AIC (AIC_c) of Hurvich and Tsai (1989). All three are based on approximations to the **Kullback-Leibler Divergence**, a fundamental quantity from information theory that is inextricably linked to maximum likelihood information.

1.1 Kullback-Leibler Divergence

Suppose that $\mathbf{y}$ is a random vector drawn from a probability distribution G with density $g(\mathbf{y})$. This is the *true DGP* and is unknown to us. Since we don't know g, we attempt to approximate it using a parametric model $f(\mathbf{y}|\theta)$, where θ is a p-vector of parameters that we estimate via maximum likelihood.[1] Since f is not the true data density, a natural question is *how well does* f

[1] I've written the model without covariates to keep the notation from getting out of control, but you could just as well write $f(\mathbf{y}|X, \theta)$. Similarly, I will sometimes write $f(\mathbf{y})$ for $f(\mathbf{y}|\theta)$ to simplify the notation below.

approximate g? It turns out that for maximum likelihood estimation there is a particularly convenient way to answer this question using the **Kullback Leibler Divergence**.

Definition 1.1 (KL Divergence)**.** *Let E_G denote expectation with respect to the true, unknown data density g. Then the Kullback-Leibler divergence from g to f is given by*

$$KL(g; f) = E_G\left[\log\left\{\frac{g(\boldsymbol{y})}{f(\boldsymbol{y})}\right\}\right] = E_G\left[\log g(\boldsymbol{y})\right] - E_G\left[\log f(\boldsymbol{y})\right]$$

The quantity $E_G\left[\log f(\mathbf{y})\right]$ is called the Expected Log-likelihood.

Key Features of the KL Divergence There are several important features to note about the KL divergence:

1. It is *not* symmetric: $KL(g; f) \neq KL(f; g)$. Hence, the KL divergence is *not* a distance function (metric).

2. $KL(g; f) \geq 0$ with equality iff $f = g$. To see why, recall that, since log is a concave function, $-\log$ is convex. Thus

$$\begin{aligned} KL(g; f) &= E_G\left[\log\left\{\frac{g(\mathbf{y})}{f(\mathbf{y})}\right\}\right] = E_G\left[-\log\left\{\frac{f(\mathbf{y})}{g(\mathbf{y})}\right\}\right] \\ &\geq -\log\left\{E_G\left[\frac{f(\mathbf{y})}{g(\mathbf{y})}\right]\right\} = -\log\left(\int g(\mathbf{y})\frac{f(\mathbf{y})}{g(\mathbf{y})}\, d\mathbf{y}\right) \\ &= -\log\left(\int f(\mathbf{y})\, d\mathbf{y}\right) = -\log(1) = 0 \end{aligned}$$

   by Jensen's Inequality. The inequality is strict only for a non-degenerate random variable and a strictly convex function. Since $-\log$ is strictly convex, the only way to make the inequality strict is for $f(\mathbf{y})/g(\mathbf{y})$ to be degenerate. This occurs precisely when $f = g$ almost everywhere.

3. Minimizing the KL divergence $KL(g; f)$ is *equivalent* to maximizing the Expected Log-Likelihood $E_G[\log f(\mathbf{y})]$. This is because the first term in the KL divergence is a constant: it in no way depends on the model $f(\mathbf{y})$. The expected Log-likelihood enters negatively:

$$KL(g; f) = E_G\left[\log g(\mathbf{y})\right] - E_G\left[\log f(\mathbf{y})\right]$$

Thus, if we can find a way to estimate the Expected Log-likelihood, we can use the KL divergence for model selection: the larger the Expected Log-likelihood, the smaller the KL divergence, and the better the model.

4. The KL divergence equals the negative of **Boltzmann's Entropy** from Statistical Mechanics. Accordingly, it represents the *information lost* when $g(\mathbf{y})$ is encoded by $f(\mathbf{y})$.

1.2 Relationship of MLE to KL

It turns out that the KL divergence is inextricably linked to maximum likelihood estimation. To make the points a little clearer, I'll assume from now on that $\mathbf{y}$ consists of iid observations Y_t for $t = 1, \ldots, T$. This is not in fact necessary for any of the derivations that follow, but it simplifies the notation. Since the expected log likelihood is unknown, we might try to approximate it using the sample analogue

$$E_{\widehat{G}}\left[\log f(\mathbf{y}, \theta)\right] = \frac{1}{T}\sum_{t=1}^{T} \log f(Y_t, \theta) = \frac{1}{T}\ell(\theta)$$

where we have replaced G with the empirical distribution $\widehat{G}$. Now, by the Weak Law of Large Numbers for iid observations

$$\frac{1}{T}\ell(\theta) \xrightarrow{p} E_G\left[\log f(\mathbf{y}, \theta)\right]$$

Under the standard regularity conditions (see Newey and McFadden, 1994) we can strengthen this result to show that

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \frac{1}{T}\ell(\theta) \xrightarrow{p} \arg\max_{\theta \in \Theta} E_G\left[\log f(\mathbf{y}, \theta)\right]$$

Since minimizing the KL divergence is the same as maximizing the expected log-likelihood we have the following result:

Proposition 1.1. *The ML estimator $\hat{\theta}$ converges in probability to the value of θ that minimizes the KL divergence from unknown true density $g(\mathbf{y})$ to the parametric family $f(\mathbf{y}|\theta)$. When $g(\mathbf{y}) = f(\mathbf{y}|\theta)$ for some value of $\theta \in \Theta$, the divergence is minimized at zero.*

1.3 A Naïve Information Criterion

If $g(\mathbf{y})$ were known, we could choose between two parametric models $f(\mathbf{y}|\theta)$ and $h(\mathbf{y}|\gamma)$ by comparing maximized Log-likelihoods. Define

$$\begin{aligned} \theta_0 &= \arg\max_{\theta \in \Theta} E_G\left[\log f(\mathbf{y}, \theta)\right] \\ \gamma_0 &= \arg\max_{\gamma \in \Gamma} E_G\left[\log h(\mathbf{y}, \gamma)\right] \end{aligned}$$

If $E_G\left[\log f(\mathbf{y}, \theta_0)\right] > E_G\left[\log h(\mathbf{y}, \gamma_0)\right]$, then the KL divergence from $g(\mathbf{y})$ to the parametric family f_θ is smaller than that from $g(\mathbf{y})$ to h_γ. Now, we know from above that $\hat{\theta} \xrightarrow{p} \theta_0$. Further, $\frac{1}{T}\ell(\theta) \xrightarrow{p} E_G\left[\log f(\mathbf{y}, \theta)\right]$. Of course, T will be constant across models, so why not use the maximized sample likelihood $\ell(\hat{\theta})$ for model comparison? Unfortunately, $\ell(\hat{\theta})$ is a *biased estimator of the expected log likelihood* because is uses the data twice: first to estimate $\hat{\theta}$ and then directly in the sum $\sum_{t=1}^{T} \log f(Y_t, \hat{\theta})$. Because $\hat{\theta}$ was chosen to conform to the idiosyncrasies of the data at hand, $\ell(\hat{\theta})$ is overly optimistic.

We can see this as follows. Since θ_0 is the population minimizer of the KL

divergence from g to f_θ, we have

$$\begin{aligned} KL\left[g(\mathbf{y}); f(\mathbf{y},\theta)\right] &\geq KL\left[g(\mathbf{y}); f(\mathbf{y},\theta_0)\right] \\ E_G\left[\log g(\mathbf{y})\right] - E_G\left[\log f(\mathbf{y},\theta)\right] &\geq E_G\left[\log g(\mathbf{y})\right] - E_G\left[\log f(\mathbf{y},\theta_0)\right] \\ E_G\left[\log f(\mathbf{y},\theta)\right] &\leq E_G\left[\log f(\mathbf{y},\theta_0)\right] \end{aligned}$$

for all $\theta \in \Theta$. Recall that $\frac{1}{T}\ell(\theta) = E_{\widehat{G}}\left[\log f(\mathbf{y},\theta)\right]$. By the definition of the maximum likelihood estimate, $\ell(\hat{\theta}) \geq \ell(\theta_0)$. Thus,

$$E_{\widehat{G}}\left[\log f(\mathbf{y},\hat{\theta})\right] \geq E_{\widehat{G}}\left[\log f\left(\mathbf{y},\theta_0\right)\right]$$

In sample, the estimate $\hat{\theta}$ will show a higher maximized log-likelihood than the value of θ that maximizes the population log-likelihood. Thus, the sample analogue is *overly optimistic*.

2 The AIC and TIC

In the previous section we explained that using the KL divergence to do model selection is equivalent to maximizing the expected log-likelihood across models. Unfortunately, using the maximized log-likelihood, based on the estimated parameters, is a biased estimator of this quantity: it is systematically too high. Both the AIC and the TIC address this problem by using asymptotic theory to get an approximate expression for the bias so that we can correct it.

To keep notation simple, throughout this section we'll assume that we have an iid sample of scalar random variables $Y_1, \ldots, Y_T$ drawn from a true but unknown distribution with density $g(y)$. As above we'll consider maximum likelihood estimation based on an approximating parametric density $f(y|\theta)$.

2.1 Fundamental Expansion for MLE

Under standard regularity conditions, see for example Newey and McFadden (1994), the maximum likelihood estimator $\widehat{\theta}$ can be expanded as

$$\widehat{\theta} = \theta_0 + J^{-1}\bar{U}_T + o_p(T^{-1/2})$$

where θ_0 is value of θ that minimizes KL divergence from g to the parametric family of distributions $f(y|\theta)$ and

$$\begin{aligned} J &= -E_G\left[\frac{\partial^2 \log f(Y|\theta)}{\partial\theta\partial\theta'}\right] \\ \bar{U}_T &= \frac{1}{T}\sum_{t=1}^{T}\frac{\partial \log f(Y_t|\theta_0)}{\partial\theta} \end{aligned}$$

Now, by the CLT, $\sqrt{T}\,\bar{U}_T \xrightarrow{d} U$ where $U \sim N_p(0, K)$ and

$$K = Var_G\left[\frac{\partial \log f(Y|\theta_0)}{\partial\theta}\right] = E_G\left[\frac{\partial \log f(Y|\theta_0)}{\partial\theta}\frac{\partial \log f(Y|\theta_0)}{\partial\theta'}\right]$$

Hence,

$$\begin{aligned} \sqrt{T}\left(\widehat{\theta} - \theta_0\right) &= \sqrt{T}\,J^{-1}\bar{U}_T + o_p(1) \\ &\xrightarrow{d} J^{-1}U \\ &\sim N_p(0, J^{-1}KJ^{-1}) \end{aligned}$$

Note that when $g = f_\theta$ for some θ, we have $K = J$ by the information matrix equality so the variance simplifies to J^{-1}.

2.2 Estimating the Expected Log Likelihood

To carry out model selection based on the KL divergence, we need to estimate the expected log likelihood. Under the iid assumption,

$$E_G[\log f(\mathbf{y}|\theta_0)] = E_G\left[\sum_{t=1}^{T} f(Y|\theta_0)\right] = T\ E_G[\log f(Y|\theta_0)]$$

so it is sufficient to work with the expected log likelihood of a single representative observation Y. Written as an integral,

$$E_G[\log f(Y|\theta_0)] = \int g(y) \log f(y|\theta_0)\ dy$$

There are two problems. First, we don't know θ_0. Of course we do have an estimator $\widehat{\theta}$, so we might consider simply plugging it in to yield

$$\int g(y) \log f(y|\theta_0)\ dy \approx \int g(y) \log f(y|\widehat{\theta})\ dy$$

Even with this approximation, however, we still don't know g, the true data density. As discussed above, trying to replace this integral with the sample analogue $\ell_T(\widehat{\theta})/T$, the maximized log-likelihood, introduces a bias. So what can we do? The idea behind the AIC and TIC is to *estimate* this bias, which we'll write relative to the infeasible plug-in estimator. In other words:

$$Bias = \frac{\ell_T(\widehat{\theta})}{T} - \int g(y) \log f(y|\widehat{\theta}) dy$$

Now, as it turns out, we can expand the bias expression as follow:

$$Bias = \bar{Z}_T + (\widehat{\theta} - \theta_0)' J (\widehat{\theta} - \theta_0) + o_p(T^{-1})$$

where

$$\bar{Z}_T = \frac{1}{T} \sum_{t=1}^{T} \{\log f(Y_t|\theta_0) - E_G[\log f(Y|\theta_0)]\}$$

You'll prove this on the problem set!

Now, recall that the bias expression depends on $\widehat{\theta}$ which is a random variable since it depends on the sample data. To address this, we will attempt to approximate the *expectation* of the bias term, where, again, the expectation is taken over the sampling distribution of $\widehat{\theta}$. Using our asymptotic expansion:

$$E[Bias] \approx E[\bar{Z}_T] + E[(\widehat{\theta} - \theta_0)' J (\widehat{\theta} - \theta_0)]$$

Since $E[\bar{Z}_T] = 0$, this becomes

$$E[Bias] \approx E[(\widehat{\theta} - \theta_0)' J (\widehat{\theta} - \theta_0)]$$

Now, using the fundamental expansion for MLE from above

$$\sqrt{T}\left(\widehat{\theta} - \theta_0\right) \xrightarrow{d} J^{-1}U$$

hence

$$T\left(\widehat{\theta} - \theta_0\right)' J \left(\widehat{\theta} - \theta_0\right) \xrightarrow{d} U' J^{-1} U$$

which suggests the approximation

$$E[Bias] \approx T^{-1} E[U' J^{-1} U]$$

Finally, using the almost magical properties of the trace operator, we have

$$\begin{aligned} E[U'J^{-1}U] &= E\left[\text{trace}\left\{U'J^{-1}U\right\}\right] = E\left[\text{trace}\left\{J^{-1}UU'\right\}\right] \\ &= \text{trace}\left\{E[J^{-1}UU']\right\} = \text{trace}\left\{J^{-1}E[UU']\right\} \\ &= \text{trace}\left\{J^{-1}K\right\} \end{aligned}$$

Thus, we approximate the expected bias by $T^{-1}\text{trace}\left\{J^{-1}K\right\}$. Finally, we correct the bias of the maximized log-likelihood and approximate the expected

log likelihood by

$$E_G[\log f(Y|\theta_0)] \approx \frac{\ell(\widehat{\theta})}{T} - \frac{\text{trace}\{J^{-1}K\}}{T}$$

multiplying through by $2/T$ and substituting consistent estimators of the matrices J and K yields **Takeuchi's Information Criterion** (TIC)

$$TIC = 2\left[\ell(\widehat{\theta}) - \text{trace}\left\{\widehat{J}^{-1}\widehat{K}\right\}\right]$$

The scaling is, of course, arbitrary but this particular choice is traditional. If there is a $\theta \in \Theta$ such that $g(y) = f(y|\theta)$ then the information matrix equality holds and $J^{-1} = K$. In this case $\text{trace}\{J^{-1}K\} = \text{trace}\{\mathbf{I}_p\} = p$. Using this quantity as the bias correction yields **Akaike's Information Criterion**

$$AIC = 2\left[\ell(\widehat{\theta}) - p\right]$$

Although the TIC and AIC are similar, there are several subtleties:

1. The bias correction for the AIC is derived under the assumption that the approximating model is *correctly specified*, while the TIC is not. In this sense the AIC is a special case of the TIC.

2. It has been argued that for models where the Information Matrix is not satisfied, the AIC will still be close to the TIC. (The log-likelihood term should dominate the bias correction in such situations.)

3. Typically, the matrices K and J are large, meaning that the estimates will have high variance (we need to estimate p^2+p elements). In contrast, the AIC a has much smaller variance because the bias correction *does not depend on the data*. Thus, even if the model is mis-specified, it may be preferable to use AIC rather than TIC unless the sample size is large.

2.3 An Important Caveat

To derive the TIC and AIC, we used the following expansion for the bias term

$$Bias = \bar{Z}_T + (\widehat{\theta} - \theta_0)' J (\widehat{\theta} - \theta_0) + o_p(T^{-1})$$

This holds under standard regularity conditions. (For details on its derivation, see the next subsection.) However, we employed a bit of sleight of hand when we proceeded to approximate the expected bias using the mean of the limiting random variable $U' J^{-1} U$. For example, the expectation of "truth" relative to which the bias is calculated, namely

$$E_G \left[\int g(y) \log f(y|\widehat{\theta}) \, dy \right]$$

does not exist in all cases. The bias correction remains reasonable in this case, as we see from the asymptotic expansion, but strictly speaking it doesn't make sense to talk about equating means.