Various Model Selection Criteria

Francis J. DiTraglia

February 26, 2014

1 Bayesian Information Criterion (BIC)

As in the derivation of AIC, simplify by looking only at a scalar random variable and ignoring dependence etc. Results are still generally true but this simplifies the notation.

1.1 Overview of the BIC

Despite its name, the BIC is *not* a Bayesian procedure. It is a large-sample Frequentist *approximation* to Bayesian model selection:

1. Begin with a uniform prior on the set of candidate models so that it suffices to maximize the Marginal Likelihood.
2. The BIC is a large sample approximation to the Marginal Likelihood:
$$\int \pi(\theta_i) f_i(\mathbf{y}|\theta_i) d\theta_i$$
3. As usual when Bayesian procedures are subjected to Frequentist asymptotics, the priors on parameters vanish in the limit.
4. We proceed by a *Laplace Approximation* to the Marginal Likelihood

1.2 Laplace Approximation

For the moment simplify the notation by suppressing dependence on M_i. We want to approximate:
$$\int \pi(\theta) f(\mathbf{y}|\theta) d\theta$$
This is actually a common problem in applications of Bayesian inference:

- Notice that $\pi(\theta)f(\mathbf{y}|\theta)$ is the *kernel* of some probability density, i.e. the density without its normalizing constant.
- *How do we know this?* By Bayes' Rule

$$\pi(\theta|\mathbf{y}) = \frac{\pi(\theta)f(\mathbf{y}|\theta)}{\int \pi(\theta)f(\mathbf{y}|\theta)d\theta}$$

  is a proper probability density and the denominator is *constant* with respect to θ. (The parameter has been "integrated out.")
- In Bayesian inference, we specify $\pi(\theta)$ and $f(\mathbf{y}|\theta)$, so $\pi(\theta)f(\mathbf{y}|\theta)$ is known. But to calculate the posterior we need to *integrate* to find the normalizing constant.
- Only in special cases (e.g. conjugate families) can we find the exact normalizing constant. Typically some kind of approximation is needed:
  - Importance Sampling
  - Markov-Chain Monte Carlo (MCMC)
  - *Laplace Approximation*

The Laplace Approximation is an *analytical approximation* based on Taylor Expansion arguments. In Bayesian applications, the expansion is carried out around the posterior mode, i.e. the mode of $\pi(\theta)f(\mathbf{y}|\theta)$, but we will expand around the Maximum likelihood estimator.

Proposition 1.1 (Laplace Approximation)**.**

$$\int \pi(\theta)f(\mathbf{y}|\theta)d\theta \approx \frac{\exp\left\{\ell(\hat{\theta})\right\}\pi(\hat{\theta})(2\pi)^{p/2}}{n^{p/2}\left|J(\hat{\theta})\right|^{1/2}}$$

Where $\hat{\theta}$ is the maximum likelihood estimator*, p the dimension of θ and*

$$J(\hat{\theta}) = -\frac{1}{n}\frac{\partial^2 \log f(\mathbf{y}|\hat{\theta})}{\partial\theta\partial\theta'}$$

Proof. A rigorous proof of this result is complicated. The following is a sketch. First write $\ell(\theta)$ for $\log f(\mathbf{y}|\theta)$ so that

$$\pi(\theta)f(\mathbf{y}|\theta) = \pi(\theta)\exp\{\log f(\mathbf{y}|\theta)\} = \pi(\theta)\exp\{\log \ell(\theta)\}$$

By a second-order Taylor Expansion around the MLE $\hat{\theta}$

$$\ell(\theta) = \ell(\hat{\theta}) + \frac{1}{2}\left(\theta - \hat{\theta}\right)' \frac{\partial^2 \ell(\hat{\theta})}{\partial\theta\partial\theta'}\left(\theta - \hat{\theta}\right) + R_\ell \tag{1}$$

since the derivative of $\ell(\theta)$ is zero at $\hat{\theta}$ by the definition of MLE. A first-order expansion is sufficient for $\pi(\theta)$ because the derivative does not vanish at $\hat{\theta}$

$$\pi(\theta) = \pi(\hat{\theta}) + \frac{\partial \pi(\hat{\theta})}{\partial\theta'}\left(\theta - \hat{\theta}\right) + R_\pi \tag{2}$$

Substituting Equations 1 and 2,

$$\begin{aligned}
\int \pi(\theta) f(\mathbf{y}|\theta) d\theta &= \int \exp\left\{\ell(\hat{\theta}) + \frac{1}{2}\left(\theta - \hat{\theta}\right)' \frac{\partial^2 \ell(\hat{\theta})}{\partial\theta\partial\theta'}\left(\theta - \hat{\theta}\right) + R_\ell\right\} \\
&\quad \times \left[\pi(\hat{\theta}) + \left(\theta - \hat{\theta}\right)' \frac{\partial \pi(\hat{\theta})}{\partial\theta} + R_\pi\right] d\theta \\
&= \exp\left\{\ell(\hat{\theta})\right\}(I_1 + I_2 + I_3)
\end{aligned}$$

where

$$\begin{aligned}
I_1 &= \pi(\hat{\theta}) \int \exp\left\{\frac{1}{2}\left(\theta - \hat{\theta}\right)' \frac{\partial^2 \ell(\hat{\theta})}{\partial\theta\partial\theta'}\left(\theta - \hat{\theta}\right) + R_\ell\right\} d\theta \\
I_2 &= \frac{\partial \pi(\hat{\theta})}{\partial\theta'} \int \left(\theta - \hat{\theta}\right) \exp\left\{\frac{1}{2}\left(\theta - \hat{\theta}\right)' \frac{\partial^2 \ell(\hat{\theta})}{\partial\theta\partial\theta'}\left(\theta - \hat{\theta}\right) + R_\ell\right\} d\theta \\
I_3 &= \int R_\pi \exp\left\{\frac{1}{2}\left(\theta - \hat{\theta}\right)' \frac{\partial^2 \ell(\hat{\theta})}{\partial\theta\partial\theta'}\left(\theta - \hat{\theta}\right) + R_\ell\right\} d\theta
\end{aligned}$$

Under certain regularity conditions (not the standard ones!) we can treat R_ℓ and R_π as approximately equal to zero for large n uniformly in θ, so that

$$\begin{aligned}
I_1 &\approx \pi(\hat{\theta}) \int \exp\left\{\frac{1}{2}\left(\theta - \hat{\theta}\right)' \frac{\partial^2 \ell(\hat{\theta})}{\partial\theta\partial\theta'}\left(\theta - \hat{\theta}\right)\right\} d\theta \\
I_2 &\approx \frac{\partial \pi(\hat{\theta})}{\partial\theta'} \int \left(\theta - \hat{\theta}\right) \exp\left\{\frac{1}{2}\left(\theta - \hat{\theta}\right)' \frac{\partial^2 \ell(\hat{\theta})}{\partial\theta\partial\theta'}\left(\theta - \hat{\theta}\right)\right\} d\theta \\
I_3 &\approx 0
\end{aligned}$$

Because $\hat{\theta}$ is the MLE,

$$\frac{\partial^2 \ell(\hat{\theta})}{\partial\theta\partial\theta'}$$

must be negative definite, so

$$-\frac{\partial^2 \ell(\hat{\theta})}{\partial\theta\partial\theta'}$$

is positive definite. It follows that

$$\exp\left\{\frac{1}{2}\left(\theta-\hat{\theta}\right)'\frac{\partial^2 \ell(\hat{\theta})}{\partial\theta\partial\theta'}\left(\theta-\hat{\theta}\right)\right\} = \exp\left\{-\frac{1}{2}\left(\theta-\hat{\theta}\right)'\left[\left(-\frac{\partial^2 \ell(\hat{\theta})}{\partial\theta\partial\theta'}\right)^{-1}\right]^{-1}\left(\theta-\hat{\theta}\right)\right\}$$

can be viewed as the kernel of a Normal distribution with mean $\hat{\theta}$ and variance matrix

$$\left(-\frac{\partial^2 \ell(\hat{\theta})}{\partial\theta\partial\theta'}\right)^{-1}$$

Thus,

$$\int \exp\left\{\frac{1}{2}\left(\theta-\hat{\theta}\right)'\frac{\partial^2 \ell(\hat{\theta})}{\partial\theta\partial\theta'}\left(\theta-\hat{\theta}\right)\right\}d\theta = (2\pi)^{p/2}\left|\left(-\frac{\partial^2 \ell(\hat{\theta})}{\partial\theta\partial\theta'}\right)^{-1}\right|^{1/2}$$

and

$$\int \left(\theta-\hat{\theta}\right)\exp\left\{\frac{1}{2}\left(\theta-\hat{\theta}\right)'\frac{\partial^2 \ell(\hat{\theta})}{\partial\theta\partial\theta'}\left(\theta-\hat{\theta}\right)\right\}d\theta = 0$$

Therefore,

$$\begin{aligned}
\int \pi(\theta)f(\mathbf{y}|\theta)d\theta &\approx \exp\left\{\ell(\hat{\theta})\right\}\pi(\hat{\theta})\,(2\pi)^{p/2}\left|\left(-\frac{\partial^2 \ell(\hat{\theta})}{\partial\theta\partial\theta'}\right)^{-1}\right|^{1/2} \\
&= \exp\left\{\ell(\hat{\theta})\right\}\pi(\hat{\theta})\,(2\pi)^{p/2}\left|n\left(-\frac{1}{n}\frac{\partial^2 \ell(\hat{\theta})}{\partial\theta\partial\theta'}\right)\right|^{-1/2} \\
&= \frac{\exp\left\{\ell(\hat{\theta})\right\}\pi(\hat{\theta})\,(2\pi)^{p/2}}{n^{p/2}\left|J(\hat{\theta})\right|^{1/2}}
\end{aligned}$$

$\square$

1.3 Finally the BIC

Now we re-introduce the dependence on the model M_i. Taking logs of the Laplace Approximation and multiplying by two (again, this is traditional but has no effect on model comparisons)

$$\begin{aligned}
2\log f(y|M_i) &= 2\log\left\{\int f_i(y|\theta_i)\pi(\theta_i)d\theta_i\right\} \\
&\approx 2\ell(\hat{\theta}_i) - p\log(n) + p\log(2\pi) - \pi(\hat{\theta}_i) - \log\left|J(\hat{\theta}_i)\right|
\end{aligned}$$

The first two terms are $O_p(n)$ and $O_p(\log n)$, while the last three are $O_p(1)$, hence negligible as $n \to \infty$. This gives us Schwarz's BIC

$$BIC(M_i) = 2\log f_i(\mathbf{y}|\hat{\theta}_i) - p\log n$$

We choose the model M_i for which $BIC(M_i)$ is largest. Notice that the prior on the parameter, $\pi(\theta)$, drops out in the limit, and recall that we began by putting a uniform prior on the *models* under consideration.

2 Hannan-Quinn

3 Final Prediction Error

4 Mallow's C_p

5 Cross-Validation

Talk about time series version and Racine paper.

6 Bootstrap Model Selection

There's a state-space paper here as well. Need to talk about block bootstrap. Mention that we're going to learn more about this when we look at Bagging later in the semester.

7 Some Time Series Examples

Reference McQuarrie and Tsai among others. Also the paper by Ng and Renault.