171:290 Model Selection
Lecture V: The Bayesian Information Criterion

Joseph E. Cavanaugh

Department of Biostatistics
Department of Statistics and Actuarial Science
The University of Iowa

September 25, 2012

Introduction

- BIC, the *Bayesian information criterion*, was introduced by Schwarz (1978) as a competitor to the Akaike (1973, 1974) information criterion.
- Schwarz derived BIC to serve as an asymptotic approximation to a transformation of the Bayesian posterior probability of a candidate model.
- In large-sample settings, the fitted model favored by BIC ideally corresponds to the candidate model which is *a posteriori* most probable; i.e., the model which is rendered most plausible by the data at hand.
- The computation of BIC is based on the empirical log-likelihood and does not require the specification of priors.

Introduction

- In Bayesian applications, pairwise comparisons between models are often based on Bayes factors.
- Assuming two candidate models are regarded as equally probable *a priori*, a Bayes factor represents the ratio of the posterior probabilities of the models. The model which is *a posteriori* most probable is determined by whether the Bayes factor is less than or greater than one.
- In certain settings, model selection based on BIC is roughly equivalent to model selection based on Bayes factors (Kass and Raftery, 1995; Kass and Wasserman, 1995).
- Thus, BIC has appeal in many Bayesian modeling problems where priors are hard to set precisely.

Introduction

Outline:

- Overview of BIC
- Derivation of BIC
- BIC and Bayes Factors
- BIC versus AIC
- Use of BIC

Overview of BIC

Key Constructs:

- **True** or **generating model**: $g(y)$.
- **Candidate** or **approximating model**: $f(y|\,\theta_k)$.
- **Candidate class:**

$$\mathcal{F}(k) = \{f(y|\,\theta_k) \mid \theta_k \in \Theta(k)\}\,.$$

- **Fitted model:** $f(y|\,\hat{\theta}_k)$.

Overview of BIC

- **Akaike information criterion:**

$$\text{AIC} = -2 \ln f(y|\hat{\theta}_k) + 2k.$$

- **Bayesian (Schwarz) information criterion:**

$$\text{BIC} = -2 \ln f(y|\hat{\theta}_k) + k \ln n.$$

- AIC and BIC feature the same goodness-of-fit term.
- The penalty term of BIC is more stringent than the penalty term of AIC. (For $n \geq 8$, $k \ln n$ exceeds $2k$.)
- Consequently, BIC tends to favor smaller models than AIC.

Overview of BIC

- The *Bayesian information criterion* is often called the *Schwarz information criterion*.
- Common acronyms: BIC, SIC, SBC, SC.
- AIC provides an asymptotically unbiased estimator of the expected Kullback discrepancy between the generating model and the fitted approximating model.
- BIC provides a large-sample estimator of a transformation of the Bayesian posterior probability associated with the approximating model.
- By choosing the fitted candidate model corresponding to the minimum value of BIC, one is attempting to select the candidate model corresponding to the highest Bayesian posterior probability.

Overview of BIC

- BIC was justified by Schwarz (1978) "for the case of independent, identically distributed observations, and linear models," under the assumption that the likelihood is from the regular exponential family.
- Generalizations of Schwarz's derivation are presented by Stone (1979), Leonard (1982), Kashyap (1982), Haughton (1988), and Cavanaugh and Neath (1999).
- We will consider a justification which is general, yet informal.

Derivation of BIC

- Let y denote the observed data.
- Assume that y is to be described using a model M_k selected from a set of candidate models $M_{k_1}, M_{k_2}, \ldots, M_{k_L}$.
- Assume that each M_k is uniquely parameterized by a vector θ_k, where θ_k is an element of the parameter space $\Theta(k)$ $(k \in \{k_1, k_2, \ldots, k_L\})$.
- Let $L(\theta_k \,|\, y)$ denote the likelihood for y based on M_k.
- Note: $L(\theta_k \,|\, y) = f(y|\, \theta_k)$.
- Let $\hat{\theta}_k$ denote the maximum likelihood estimate of θ_k obtained by maximizing $L(\theta_k \,|\, y)$ over $\Theta(k)$.

Derivation of BIC

- We assume that derivatives of $L(\theta_k \mid y)$ up to order two exist with respect to θ_k, and are continuous and suitably bounded for all $\theta_k \in \Theta(k)$.
- The motivation behind BIC can be seen through a Bayesian development of the model selection problem.
- Let $\pi(k)$ $(k \in \{k_1, k_2, \ldots, k_L\})$ denote a discrete prior over the models $M_{k_1}, M_{k_2}, \ldots, M_{k_L}$.
- Let $g(\theta_k \mid k)$ denote a prior on θ_k given the model M_k $(k \in \{k_1, k_2, \ldots, k_L\})$.

Derivation of BIC

- Applying Bayes' Theorem, the joint posterior of M_k and θ_k can be written as

$$h((k,\ \theta_k)\,|\,y) = \frac{\pi(k)\ g(\theta_k\,|\,k)\ L(\theta_k\,|\,y)}{m(y)},$$

  where $m(y)$ denotes the marginal distribution of y.
- A Bayesian model selection rule might aim to choose the model M_k which is *a posteriori* most probable.
- The posterior probability for M_k is given by

$$P(k\,|\,y) = m(y)^{-1}\ \pi(k) \int L(\theta_k\,|\,y)\ g(\theta_k\,|\,k)\ d\theta_k.$$

Derivation of BIC

- Now consider minimizing $-2 \ln P(k \mid y)$ as opposed to maximizing $P(k \mid y)$.
- We have

$$\begin{aligned} -2 \ln P(k \mid y) &= 2 \ln \{m(y)\} - 2 \ln \{\pi(k)\} \\ &\quad -2 \ln \left\{ \int L(\theta_k \mid y) \; g(\theta_k \mid k) \; d\theta_k \right\}. \end{aligned}$$

- The term involving $m(y)$ is constant with respect to k; thus, for the purpose of model selection, this term can be discarded.

Derivation of BIC

- We obtain

$$\begin{aligned} -2\ln P(k \mid y) &\propto -2\ln\{\pi(k)\} \\ &\quad -2\ln\left\{\int L(\theta_k \mid y)\; g(\theta_k \mid k)\; d\theta_k\right\} \\ &\equiv S(k \mid y). \end{aligned}$$

- Now consider the integral which appears above:

$$\int L(\theta_k \mid y)\; g(\theta_k \mid k)\; d\theta_k.$$

- In order to obtain an approximation to this term, we take a second-order Taylor series expansion of the log-likelihood about $\hat{\theta}_k$.

Derivation of BIC

We have

$$\begin{aligned}
\ln L(\theta_k \,|\, y) &\approx \ln L(\hat{\theta}_k \,|\, y) + (\theta_k - \hat{\theta}_k)' \, \frac{\partial \ln L(\hat{\theta}_k \,|\, y)}{\partial \theta_k} \\
&\quad + \frac{1}{2}(\theta_k - \hat{\theta}_k)' \left[\frac{\partial^2 \ln L(\hat{\theta}_k \,|\, y)}{\partial \theta_k \partial \theta_k'} \right] (\theta_k - \hat{\theta}_k) \\
&= \ln L(\hat{\theta}_k \,|\, y) - \frac{1}{2}(\theta_k - \hat{\theta}_k)' \left[n \, \bar{\mathcal{I}}(\hat{\theta}_k, y) \right] (\theta_k - \hat{\theta}_k),
\end{aligned}$$

where

$$\bar{\mathcal{I}}(\hat{\theta}_k, y) = -\frac{1}{n} \, \frac{\partial^2 \ln L(\hat{\theta}_k \,|\, y)}{\partial \theta_k \partial \theta_k'}$$

is the *average* observed Fisher information matrix.

Derivation of BIC

- Thus,

$$L(\theta_k \,|\, y) \approx L(\hat{\theta}_k \,|\, y) \, \exp\left\{ -\frac{1}{2} \, (\theta_k - \hat{\theta}_k)' \, \left[n \, \bar{\mathcal{I}}(\hat{\theta}_k, y) \right] \, (\theta_k - \hat{\theta}_k) \right\}.$$

- We therefore have the following approximation for our integral:

$$\int L(\theta_k \,|\, y) \, g(\theta_k \,|\, k) \, d\theta_k \approx$$

$$L(\hat{\theta}_k \,|\, y) \int \exp\left\{ -\frac{1}{2} \, (\theta_k - \hat{\theta}_k)' \, \left[n \, \bar{\mathcal{I}}(\hat{\theta}_k, y) \right] \, (\theta_k - \hat{\theta}_k) \right\} g(\theta_k \,|\, k) \, d\theta_k.$$

Derivation of BIC

- Now consider the evaluation of

$$\int \exp\left\{-\frac{1}{2}\,(\theta_k-\hat{\theta}_k)'\,\left[n\,\bar{\mathcal{I}}(\hat{\theta}_k,y)\right]\,(\theta_k-\hat{\theta}_k)\right\}\,g(\theta_k\,|\,k)\,d\theta_k$$

  using the noninformative prior $g(\theta_k\,|\,k)\;=\;1$.

- In this case, we obtain

$$\int \exp\left\{-\frac{1}{2}\,(\theta_k-\hat{\theta}_k)'\,\left[n\,\bar{\mathcal{I}}(\hat{\theta}_k,y)\right]\,(\theta_k-\hat{\theta}_k)\right\}\,d\theta_k = (2\pi)^{(k/2)}\,|n\,\bar{\mathcal{I}}(\hat{\theta}_k,y)|^{-1/2}.$$

Derivation of BIC

- We therefore have

$$\begin{aligned}
&\int L(\theta_k \,|\, y)\; g(\theta_k \,|\, k)\; d\theta_k \\
&\approx L(\hat{\theta}_k \,|\, y)\; (2\pi)^{(k/2)}\; |n\, \bar{\mathcal{I}}(\hat{\theta}_k, y)|^{-1/2} \\
&= L(\hat{\theta}_k \,|\, y)\; (2\pi)^{(k/2)}\; n^{(-k/2)}\; |\bar{\mathcal{I}}(\hat{\theta}_k, y)|^{-1/2} \\
&= L(\hat{\theta}_k \,|\, y)\; \left(\frac{2\pi}{n}\right)^{(k/2)}\; |\bar{\mathcal{I}}(\hat{\theta}_k, y)|^{-1/2}.
\end{aligned}$$

Derivation of BIC

- The preceding can be viewed as a variation on the Laplace method of approximating the integral

$$\int L(\theta_k \,|\, y)\; g(\theta_k \,|\, k)\; d\theta_k.$$

  (See Tierney and Kadane, 1986; Kass and Raftery, 1995.)
- This approximation is valid in large-sample settings provided that the prior $g(\theta_k \,|\, k)$ is "flat" over the neighborhood of $\hat{\theta}_k$ where $L(\theta_k \,|\, y)$ is dominant.
- The prior $g(\theta_k \,|\, k)$ need not be noninformative, although the choice of $g(\theta_k \,|\, k) \;=\; 1$ makes our derivation more tractable.

Derivation of BIC

- We can now write

$$
\begin{aligned}
S(k \mid y) &= -2 \ln\{\pi(k)\} \\
&\quad -2 \ln \left\{ \int L(\theta_k \mid y)\ g(\theta_k \mid k)\ d\theta_k \right\} \\
&\approx -2 \ln\{\pi(k)\} \\
&\quad -2 \ln \left[L(\hat{\theta}_k \mid y)\ \left(\frac{2\pi}{n} \right)^{(k/2)}\ |\bar{\mathcal{I}}(\hat{\theta}_k, y)|^{-1/2} \right] \\
&= -2 \ln\{\pi(k)\} \\
&\quad -2 \ln L(\hat{\theta}_k \mid y) + k \left\{ \ln \left(\frac{n}{2\pi} \right) \right\} + \ln |\bar{\mathcal{I}}(\hat{\theta}_k, y)|.
\end{aligned}
$$

Derivation of BIC

- Ignoring terms in the preceding that are bounded as the sample size grows to infinity, we obtain

$$S(k \mid y) \approx -2 \ln L(\hat{\theta}_k \mid y) + k \ln n.$$

- With this motivation, the Bayesian (Schwarz) information criterion is defined as follows:

$$\begin{aligned} \text{BIC} &= -2 \ln L(\hat{\theta}_k \mid y) + k \ln n \\ &= -2 \ln f(y \mid \hat{\theta}_k) + k \ln n. \end{aligned}$$

BIC and Bayes Factors

- Consider two candidate models M_{k_1} and M_{k_2} in a Bayesian analysis. To choose between these models, a *Bayes factor* is often used.
- The **Bayes factor**, B_{12}, is defined as a ratio of the posterior odds of M_{k_1},

$$P(k_1 \,|\, y)/P(k_2 \,|\, y),$$

to the prior odds of M_{k_1},

$$\pi(k_1)/\pi(k_2).$$

- If $B_{12} > 1$, model M_{k_1} is favored by the data; if $B_{12} < 1$, model M_{k_2} is favored by the data.

BIC and Bayes Factors

- Kass and Raftery (1995) write "The Bayes factor is a summary of the evidence provided by the data in favor of one scientific theory, represented by a statistical model, as opposed to another."

BIC and Bayes Factors

- Let $\text{BIC}(k_1)$ denote BIC for model M_{k_1}, and let $\text{BIC}(k_2)$ denote BIC for model M_{k_2}. Kass and Raftery (1995) argue that as $n \to \infty$,

$$\frac{-2 \ln B_{12} - (\text{BIC}(k_1) - \text{BIC}(k_2))}{-2 \ln B_{12}} \to 0.$$

- Thus, $(\text{BIC}(k_1) - \text{BIC}(k_2))$ can be viewed as a rough approximation to $-2 \ln B_{12}$.
- Kass and Raftery (1995) write "The Schwarz criterion (or BIC) gives a rough approximation to [-2] the logarithm of the Bayes factor, which is easy to use and does not require evaluation of prior distributions. It is well suited for summarizing results in scientific communication."

BIC versus AIC

- Recall the definitions of consistency and asymptotic efficiency.
- Suppose that the generating model is of a finite dimension, and that this model is represented in the candidate collection under consideration. A **consistent** criterion will asymptotically select the fitted candidate model having the correct structure with probability one.
- On the other hand, suppose that the generating model is of an infinite dimension, and therefore lies outside of the candidate collection under consideration. An **asymptotically efficient** criterion will asymptotically select the fitted candidate model which minimizes the mean squared error of prediction.
- AIC is asymptotically efficient yet not consistent;
  BIC is consistent yet not asymptotically efficient.

BIC versus AIC

- AIC and BIC share the same goodness-of-fit term, but the penalty term of BIC ($k \ln n$) is potentially much more stringent than the penalty term of AIC ($2k$).
- Thus, BIC tends to choose fitted models that are more parsimonious than those favored by AIC.
- The differences in selected models may be especially pronounced in large sample settings.
- Intuitively, why is the complexity penalization so much greater for BIC than for AIC?

BIC versus AIC

- In Bayesian analyses, the strength of evidence required to favor additional complexity is generally greater than in frequentist analyses. Why?
  - The Bayesian analytical paradigm incorporates *estimation uncertainty* and *parameter uncertainty*.
  - The frequentist analytical paradigm only incorporates *estimation uncertainty*.

BIC versus AIC

From a practical perspective,

- AIC could be advocated when the primary goal of the modeling application is *predictive*; i.e., to build a model that will effectively predict new outcomes.
- BIC could be advocated when the primary goal of the modeling application is *descriptive*; i.e., to build a model that will feature the most meaningful factors influencing the outcome, based on an assessment of relative importance.
- As the sample size grows, predictive accuracy improves as subtle effects are admitted to the model. AIC will increasingly favor the inclusion of such effects; BIC will not.

Use of BIC

- BIC can be used to compare non-nested models.
- BIC can be used to compare models based on different probability distributions. However, when the criterion values are computed, no constants should be discarded from the goodness-of-fit term $-2 \ln f(y|\hat{\theta}_k)$.

Use of BIC

- In a model selection application, the optimal fitted model is identified by the minimum value of BIC.
- However, as with the application of any model selection criterion, the criterion values are important; models with similar values should receive the same "ranking" in assessing criterion preferences.

Use of BIC

- Question: What constitutes a substantial difference in criterion values?
- For BIC, Kass and Raftery (1995, p. 777) feature the following table (slightly revised for presentation).

$\text{BIC}_i - \text{BIC}_{min}$	Evidence Against Model i
0 - 2	Not worth more than a bare mention
2 - 6	Positive
6 - 10	Strong
> 10	Very Strong

Use of BIC

- The use of BIC seems justifiable for model screening in large-sample Bayesian analyses.
- However, BIC is often employed in frequentist analyses.
- Some frequentist practitioners prefer BIC to AIC, since BIC tends to choose fitted models that are more parsimonious than those favored by AIC.
- However, given the Bayesian justification of BIC, is the use of the criterion in frequentist analyses defensible?

References

- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* **6**, 461–464.
- Kass, R. and Raftery, A. (1995). Bayes Factors. *Journal of the American Statistical Association* **90**, 773–795.
- Neath, A. A. and Cavanaugh, J. E. (2012). The Bayesian information criterion: Background, derivation, and applications. *WIREs Computational Statistics* **4**, 199–203.