On the derivation of the Bayesian Information Criterion

H. S. Bhat[*][†] N. Kumar[*]

November 8, 2010

Abstract

We present a careful derivation of the Bayesian Inference Criterion (BIC) for model selection. The BIC is viewed here as an approximation to the Bayes Factor. One of the main ingredients in the approximation, the use of Laplace's method for approximating integrals, is explained well in the literature. Our derivation sheds light on this and other steps in the derivation, such as the use of a flat prior and the invocation of the weak law of large numbers, that are not often discussed in detail.

1 Notation

Let us define the notation that we will use:

$$
\begin{aligned}
\mathbf{y} &: \text{observed data } y_1, \ldots, y_n \\
M_i &: \text{candidate model} \\
P(\mathbf{y}|M_i) &: \text{marginal likelihood of the model } M_i \text{ given the data} \\
\boldsymbol{\theta_i} &: \text{vector of parameters in the model } M_i \\
g_i(\boldsymbol{\theta}_i) &: \text{the prior density of the parameters } \boldsymbol{\theta}_i \\
f(\mathbf{y}|\boldsymbol{\theta}_i) &: \text{the density of the data given the parameters } \boldsymbol{\theta}_i \\
L(\boldsymbol{\theta}_i|\mathbf{y}) &: \text{the likelihood of } \mathbf{y} \text{ given the model } M_i \\
\hat{\boldsymbol{\theta}}_i &: \text{the MLE of } \boldsymbol{\theta}_i \text{ that maximizes } L(\boldsymbol{\theta}_i|\mathbf{y})
\end{aligned}
$$

2 Bayes Factor

The Bayesian approach to model selection [1] is to maximize the posterior probability of a model (M_i) given the data $\{y_j\}_{j=1}^n$. Applying Bayes theorem to calculate the posterior

[*] School of Natural Sciences, University of California, Merced, 5200 N. Lake Rd., Merced, CA 95343

[†] Corresponding author, email: hbhat@ucmerced.edu

probability of a model given the data, we get

$$P(M_i|y_1,\ldots,y_n)=\frac{P(y_1,\ldots,y_n|M_i)P(M_i)}{P(y_1,\ldots,y_n)},\tag{1}$$

where $P(y_1,\ldots,y_n|M_i)$ is called the marginal likelihood of the model M_i.

If all candidate models are equally likely, then maximizing the posterior probability of a model given the data is the same as maximizing the marginal likelihood

$$P(y_1,\ldots,y_n|M_i)=\int_{\boldsymbol{\Theta}_i} L(\boldsymbol{\theta}_i|y_1,\ldots,y_n)g_i(\boldsymbol{\theta}_i)d\boldsymbol{\theta}_i,\tag{2}$$

where $\boldsymbol{\theta}_i$ denotes the vector of parameters in the model M_i, L is the likelihood function and $g_i(\boldsymbol{\theta}_i)$ is the p.d.f. of the distribution of parameters $\boldsymbol{\theta}_i$.

3 Derivation of the BIC

3.1 Laplace's Method

Let us first remind ourselves of the Laplace's method for approximating an integral.

$$\int_a^b e^{Mf(x)}\,dx\approx\sqrt{\frac{2\pi}{M|f''(x_0)|}}e^{Mf(x_0)}\text{ as }M\to\infty.$$

For this approximation to hold, the function f should have one global maximum and it should decay rapidly to zero away from the maximum.

Let us now calculate the Bayes factor $B_{01}(\mathbf{y})=P(\mathbf{y}|M_0)/P(\mathbf{y}|M_1)$ that is used to choose between two models 0 and 1. To do this, we need to calculate

$$P(\mathbf{y}|M_i)=\int f(\mathbf{y}|\boldsymbol{\theta}_i)g_i(\boldsymbol{\theta}_i)d\boldsymbol{\theta}_i=\int\exp\Big(\log\big(f(\mathbf{y}|\boldsymbol{\theta}_i)g_i(\boldsymbol{\theta}_i)\big)\Big)d\boldsymbol{\theta}_i.$$

We can now expand $\log\big(f(\mathbf{y}|\boldsymbol{\theta}_i)g_i(\boldsymbol{\theta}_i)\big)$ about its posterior mode $\tilde{\boldsymbol{\theta}}_i$ where $f(\mathbf{y}|\boldsymbol{\theta}_i)g_i(\boldsymbol{\theta}_i)$ attains its maximum and, consequently, $\log\big(f(\mathbf{y}|\boldsymbol{\theta}_i)g_i(\boldsymbol{\theta}_i)\big)$ also attains its maximum. Thus, we can approximate

$$\overbrace{\log\big(f(\mathbf{y}|\boldsymbol{\theta}_i)g_i(\boldsymbol{\theta}_i)\big)}^{Q}\approx\log\big(f(\mathbf{y}|\tilde{\boldsymbol{\theta}}_i)g_i(\tilde{\boldsymbol{\theta}}_i)\big)+(\boldsymbol{\theta}_i-\tilde{\boldsymbol{\theta}}_i)\nabla_{\boldsymbol{\theta}_i}Q|_{\tilde{\boldsymbol{\theta}}_i}+\frac{1}{2}(\boldsymbol{\theta}_i-\tilde{\boldsymbol{\theta}}_i)^T H_{\boldsymbol{\theta}_i}(\boldsymbol{\theta}_i-\tilde{\boldsymbol{\theta}}_i),$$

where $H_{\boldsymbol{\theta}_i}$ is a $|\boldsymbol{\theta}_i||\boldsymbol{\theta}_i|$ matrix such that $H_{mn}=\partial^2Q/\partial\theta_m\partial\theta_n|_{\tilde{\boldsymbol{\theta}}_i}$. Since Q attains its maximum at $\tilde{\boldsymbol{\theta}}_i$, the Hessian matrix $H_{\boldsymbol{\theta}_i}$ is negative definite. Let us denote $\tilde{H}_{\boldsymbol{\theta}_i}=-H_{\boldsymbol{\theta}_i}$, and then approximate $P(\mathbf{y}|M_i)$:

$$P(\mathbf{y}|M_i)\approx\int\exp\left\{Q|_{\tilde{\boldsymbol{\theta}}_i}+(\boldsymbol{\theta}_i-\tilde{\boldsymbol{\theta}}_i)\nabla_{\boldsymbol{\theta}_i}Q|_{\tilde{\boldsymbol{\theta}}_i}-\frac{1}{2}(\boldsymbol{\theta}_i-\tilde{\boldsymbol{\theta}}_i)^T\tilde{H}_{\boldsymbol{\theta}_i}(\boldsymbol{\theta}_i-\tilde{\boldsymbol{\theta}}_i)\right\}d\boldsymbol{\theta}_i.$$

Since Q attains its maximum at $\tilde{\boldsymbol{\theta}}_i$, we see that $\nabla_{\boldsymbol{\theta}_i} Q|_{\tilde{\boldsymbol{\theta}}_i} = 0$. Hence

$$\begin{aligned} P(\mathbf{y}|M_i) &\approx \exp(Q|_{\tilde{\boldsymbol{\theta}}_i}) \int \exp\left\{ -\frac{1}{2}(\boldsymbol{\theta}_i - \tilde{\boldsymbol{\theta}}_i)^T \tilde{H}_{\boldsymbol{\theta}_i} (\boldsymbol{\theta}_i - \tilde{\boldsymbol{\theta}}_i) \right\} d\boldsymbol{\theta}_i \\ &= \exp(Q|_{\tilde{\boldsymbol{\theta}}_i}) \int \exp\left\{ -\frac{1}{2} X^T \tilde{H}_{\boldsymbol{\theta}_i} X \right\} dX. \end{aligned}$$

Since the matrix $\tilde{H}_{\boldsymbol{\theta}_i}$ is symmetric (by virtue of being the negative of the Hessian matrix), we can diagonalize it as $\tilde{H}_{\boldsymbol{\theta}_i} = S^T \Lambda S$. Let us make a substitution $X = S^T U$ to evaluate the integral above. The Jacobian matrix $J_{mn}(U) = \partial X_m / \partial U_n \Rightarrow J(U) = S^T$. Thus $\det J(U) = 1$, and

$$\begin{aligned} P(\mathbf{y}|M_i) &\approx \exp(Q|_{\tilde{\boldsymbol{\theta}}_i}) \int \exp\left\{ -\frac{1}{2} U^T \Lambda U \right\} (\det J(U)) dU \\ &= \exp(Q|_{\tilde{\boldsymbol{\theta}}_i}) \int \exp\left\{ -\frac{1}{2} \sum_{j=1}^{|\boldsymbol{\theta_i}|} \lambda_j U_j^2 \right\} dU \\ &= \exp(Q|_{\tilde{\boldsymbol{\theta}}_i}) \prod_{j=1}^{|\boldsymbol{\theta_i}|} \sqrt{\frac{2\pi}{\lambda_j}} \\ &= \exp(Q|_{\tilde{\boldsymbol{\theta}}_i}) \frac{(2\pi)^{|\boldsymbol{\theta_i}|/2}}{\prod_j^{|\boldsymbol{\theta_i}|} \lambda_j^{1/2}} \\ &= f(y|\tilde{\boldsymbol{\theta}}_i) g_i(\tilde{\boldsymbol{\theta}}_i) \frac{2\pi^{|\boldsymbol{\theta_i}|/2}}{|\tilde{H}_{\boldsymbol{\theta}_i}|^{1/2}}, \end{aligned} \tag{3}$$

where λ_j is the j-th eigenvalue of the matrix $\tilde{H}_{\boldsymbol{\theta}_i}$. Taking log of (3), we get

$$2 \log P(\mathbf{y}|M_i) = 2 \log f(\mathbf{y}|\tilde{\boldsymbol{\theta}}_i) + 2 \log g_i(\tilde{\boldsymbol{\theta}}_i) + |\boldsymbol{\theta}_i| \log(2\pi) + \log |\tilde{H}_{\boldsymbol{\theta}_i}^{-1}|. \tag{4}$$

3.2 Flat Prior and the Weak Law of Large Numbers

Since the observed data $\mathbf{y}$ is given, $f(\mathbf{y}|\boldsymbol{\theta}_i)$ is the likelihood $L(\boldsymbol{\theta}_i|\mathbf{y})$ and L attains its maximum at the maximum likelihood estimate $\boldsymbol{\theta}_i = \hat{\boldsymbol{\theta}}_i$. If we set $g_i(\boldsymbol{\theta}_i) = 1$, an uninformative, flat prior, then each element in the matrix, $\tilde{H}_{\boldsymbol{\theta}_i}$ can be expressed as

$$\tilde{H}_{mn} = -\frac{\partial^2 \log L(\boldsymbol{\theta}_i|\mathbf{y})}{\partial \theta_m \partial \theta_n}\Big|_{\boldsymbol{\theta}_i = \hat{\boldsymbol{\theta}}_i}.$$

The matrix $\tilde{H}_{\boldsymbol{\theta}_i}$ is the observed Fisher information matrix. We can further represent every entry in the observed Fisher information matrix as

$$\begin{aligned}\tilde{H}_{mn} &= -\frac{\partial^2 \log(\prod_{j=1}^{n} L(\boldsymbol{\theta}_i|y_j))}{\partial\theta_m\partial\theta_n}\Big|_{\boldsymbol{\theta}_i=\hat{\boldsymbol{\theta}}_i} \\ &= -\frac{\partial^2 \sum_{j=1}^{n} \log L(\boldsymbol{\theta}_i|y_j)}{\partial\theta_m\partial\theta_n}\Big|_{\boldsymbol{\theta}_i=\hat{\boldsymbol{\theta}}_i} \\ &= -\frac{\partial^2 \left(\frac{1}{n}\sum_{j=1}^{n} n\log L(\boldsymbol{\theta}_i|y_j)\right)}{\partial\theta_m\partial\theta_n}\Big|_{\boldsymbol{\theta}_i=\hat{\boldsymbol{\theta}}_i}.\end{aligned}$$

At this point, we assume that the observed data $y_1, \ldots, y_n$ is IID and that n is large. This allows us to invoke the weak law of large numbers on the random variable $X_j = n\log L(\boldsymbol{\theta}_i|y_j)$. We obtain

$$\frac{1}{n}\sum_{j=1}^{n} n\log L(\boldsymbol{\theta}_i|y_j) \xrightarrow{\text{P}} E[n\log L(\boldsymbol{\theta}_i|y_j)]. \tag{5}$$

Using (5), every element in the observed Fisher information matrix is

$$\begin{aligned}\tilde{H}_{mn} &= -\frac{\partial^2 E[n\log L(\boldsymbol{\theta}_i|y_j)]}{\partial\theta_m\partial\theta_n}\Big|_{\boldsymbol{\theta}_i=\hat{\boldsymbol{\theta}}_i} \\ &= -n\frac{\partial^2 E[\log L(\boldsymbol{\theta}_i|y_j)]}{\partial\theta_m\partial\theta_n}\Big|_{\boldsymbol{\theta}_i=\hat{\boldsymbol{\theta}}_i} \\ &= -n\frac{\partial^2 E[\log L(\boldsymbol{\theta}_i|y_1)]}{\partial\theta_m\partial\theta_n}\Big|_{\boldsymbol{\theta}_i=\hat{\boldsymbol{\theta}}_i} \\ &= nI_{mn},\end{aligned}$$

so

$$|\tilde{H}_{\boldsymbol{\theta}_i}| = n^{|\boldsymbol{\theta}_i|}|I_{\boldsymbol{\theta}_i}|, \tag{6}$$

where $I_{\boldsymbol{\theta}_i}$ is the Fisher information matrix for a single data point y_1. Plugging the result from (6) into (4), we obtain

$$2\log P(\mathbf{y}|M_i) = 2\log L(\hat{\boldsymbol{\theta}}_i|\mathbf{y}) + 2\log g_i(\tilde{\boldsymbol{\theta}}_i) + |\boldsymbol{\theta}_i|\log(2\pi) - |\boldsymbol{\theta}_i|\log n - \log|I_{\boldsymbol{\theta}_i}|. \tag{7}$$

For large n, keeping the terms involving n and ignoring the rest, we find

$$\log P(\mathbf{y}|M_i) = \log L(\hat{\boldsymbol{\theta}}_i|\mathbf{y}) - \frac{|\boldsymbol{\theta}_i|}{2}\log n. \tag{8}$$

The right-hand side of (8) is the BIC estimate for a model M_i.

References

[1] J. K. Ghosh, M. Delampady, and T. Samanta, *An Introduction to Bayesian Analysis: Theory and Methods.* Springer-Verlag, New York, 2006.