A Laplace Approximation for Approximate Bayesian Model Selection

Richard M. Golden[*] **Shaurabh Nandy, Vishal Patel, Pratibha Viraktamath**
University of Texas at Dallas, BBS, GR4.1, Richardson, TX 75080
golden@utdallas.edu

Abstract

Bayesian model selection criteria (BMSC) require the evaluation of a computationally intractable multidimensional integral. Although computationally expensive Monte Carlo simulation methods may be used for such evaluations, Laplace approximation methods provide a computationally inexpensive methodology. In this paper, the computationally intractable multidimensional BMSC integral is approximated with an alternative integrand and a Laplace approximation is applied to obtain a new BMSC called "GBIC$_X$". With respect to seven real-world data sets, GBIC$_X$ exhibited performance which was superior to BIC-family model selection criteria for AIC-biased simulation studies and showed performance which was superior to AIC-family model selection criteria for BIC-biased simulation studies. These findings suggest that GBIC$_X$ may be especially useful in situations where a more robust BMSC approximation is desirable.

An important task in machine learning is the comparison of probabilistic models. One approach to model selection is Bayesian Model Selection. The critical principle of Bayesian model selection is that one computes the posterior likelihood of a model $\mathcal{M}$ given the observed data $\mathcal{D}_n$ (e.g., [1]).

Definition (Bayesian Model Selection Criterion (BMSC))**.** Let the random sample $\tilde{\mathcal{D}}_n \equiv [\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_n]$ be a sequence of $i.i.d.$ d-dimensional random vectors with common Radon-Nikodym density $p_o : \mathcal{R}^d \to [0, \infty)$ defined with respect to measure ν. Let $\mathcal{D}_n \equiv [\mathbf{x}_1, \ldots, \mathbf{x}_n]$ denote a realization of $\tilde{\mathcal{D}}_n$. Let the *parameter space* $\Theta_{\mathcal{M}}$ be a closed and bounded subset of $\mathcal{R}^q$. An element $\boldsymbol{\theta} \in \Theta_{\mathcal{M}}$ is called a *parameter vector.* Let the model specification $\mathcal{M} \equiv \{p(\cdot|\boldsymbol{\theta}, \mathcal{M}) : \boldsymbol{\theta} \in \Theta_{\mathcal{M}}\}$ with respect to measure ν. Let the *model parameter prior* $p_\theta(\cdot|\mathcal{M}) : \Theta_{\mathcal{M}} \to [0, \infty)$ be an absolutely continuous density. Let the *likelihood function*

$$p(\mathcal{D}_n|\boldsymbol{\theta}, \mathcal{M}) \equiv \prod_{i=1}^{n} p(\mathbf{x}_i|\boldsymbol{\theta}, \mathcal{M}).$$

Let $\ell_n(\boldsymbol{\theta}) \equiv -(1/n) \log p(\mathcal{D}_n|\boldsymbol{\theta}, \mathcal{M})$ be called the *negative normalized log-likelihood* (NNL) for model $\mathcal{M}$. Let the *marginal likelihood*

$$p(\mathcal{D}_n|\mathcal{M}) \equiv \int_{\Theta} p(\mathcal{D}_n|\boldsymbol{\theta}, \mathcal{M}) p_\theta(\boldsymbol{\theta}|\mathcal{M}) d\boldsymbol{\theta}. \tag{1}$$

Let the *model prior* $p_M : \{\mathcal{M}_1, \ldots, \mathcal{M}_M\} \to [0, 1]$ be a probability mass function. Let the *posterior model probability*

$$p(\mathcal{M}|\mathcal{D}_n) \equiv \frac{p_M(\mathcal{M}) p(\mathcal{D}_n|\mathcal{M})}{p(\mathcal{D}_n)}. \tag{2}$$

Then the *BMSC* (Bayesian Model Selection Criterion) for model $\mathcal{M}$ is defined as:

$$BMSC = -(1/n) \log p(\mathcal{M}|\mathcal{D}_n). \tag{3}$$

[*] www.utdallas.edu/~golden

Note that if one assumes that $p_M(\mathcal{M}_k) = (1/M)$ for $k = 1, \ldots, M$ then selecting the *most probable model* $\mathcal{M}$ given the observed data $\mathcal{D}_n$ is equivalent to finding the $\mathcal{M}_k$ which maximizes $p(\mathcal{D}_n|\mathcal{M}_k)$ for $k \in \{1, \ldots, M\}$.

Note that the multidimensional integral in (1) is typically computationally intractable and is often evaluated using Monte Carlo simulation methods (e.g., [2]). Such Monte Carlo simulation methods are computationally expensive. An alternative to such methods which can in some cases yield comparable results at a fraction of the computational expense are methods based upon the Multidimensional Laplace Approximation (e.g., [2]).

1 Bayesian Model Selection Criteria

Let the loss function $c : \mathcal{R}^d \times \Theta \rightarrow \mathcal{R}$ be defined such that $c(\mathbf{x}, \boldsymbol{\theta})$ denotes the loss incurred for experiencing event $\mathbf{x}$ when the learning machine parameter values have been set equal to $\boldsymbol{\theta}$. In this paper, it is assumed that

$$c(\mathbf{x}, \boldsymbol{\theta}) = -\log p(\mathbf{x}|\boldsymbol{\theta}, \mathcal{M})$$

so that the random *empirical risk function* $\tilde{\ell}_n(\boldsymbol{\theta}) \equiv (1/n) \sum_{i=1}^{n} c(\tilde{\mathbf{x}}_i, \boldsymbol{\theta})$ is a random *negative normalized log-likelihood function*.

Let $\tilde{\mathbf{A}}_n \equiv \nabla^2 \tilde{\ell}_n$. Let $\tilde{\mathbf{B}}_n \equiv (1/n) \sum_{i=1}^{n} \nabla c(\tilde{\mathbf{x}}_i, \cdot)[\nabla c(\tilde{\mathbf{x}}_i, \cdot)]^T$. Let $\hat{\ell}_n \equiv \tilde{\ell}_n(\hat{\boldsymbol{\theta}}_n)$.
Let $\hat{\mathbf{A}}_n \equiv \tilde{\mathbf{A}}_n(\hat{\boldsymbol{\theta}}_n)$. Let $\hat{\mathbf{B}}_n \equiv \tilde{\mathbf{B}}_n(\hat{\boldsymbol{\theta}}_n)$.

Laplace Approximation methods have been used to obtain computationally tractable approximations to the integral in (1) given regularity assumptions provided in the last section of this paper. For example, using Laplace Approximation methods, $-2 \log p(\tilde{\mathcal{D}}_n|\mathcal{M})$ may be approximated by the Bayesian Information Criterion (BIC) ([3])

$$\text{BIC} = 2n\tilde{\ell}_n(\hat{\boldsymbol{\theta}}_n) + q \log(n).$$

Higher-order Laplace approximations of $-2 \log p(\tilde{\mathcal{D}}_n|\mathcal{M})$ have also been developed. For example, a classical expansion is discussed by ([1]) and will be referred to as the Laplace Generalized Bayesian Information Criterion (GBIC$_L$). In particular, using the deterministic Laplace Approximation ([2], pp. 86-88)

$$\text{GBIC}_L = 2n\tilde{\ell}_n(\hat{\boldsymbol{\theta}}_n) - 2 \log \left[p_\theta(\hat{\boldsymbol{\theta}}_n|\mathcal{M}) \right] + q \log \left(\frac{n}{2\pi} \right) + \log(\det(\hat{\mathbf{A}}_n)).$$

More recently, Lv and Liu (2014) (see [4]) have proposed two important new high-order Laplace Approximation formulas called as GBIC and GBIC$_P$ that use assumptions about the data generating process which differ from the assumptions used in the derivation of GBIC$_L$.

Finally, the Akaike Information Criterion (AIC) ([5, 6])

$$\text{AIC} = 2n\tilde{\ell}_n(\hat{\boldsymbol{\theta}}_n) + 2q$$

and the Generalized Akaike Information Criterion (GAIC) ([6, 7])

$$\text{GAIC} = 2n\tilde{\ell}_n(\hat{\boldsymbol{\theta}}_n) + 2TRACE\left(\hat{\mathbf{A}}_n^{-1} \hat{\mathbf{B}}_n \right)$$

are asymptotically equivalent to cross-validation estimators of the expected NNL (i.e., $2nE\{\hat{l}_n\}$) (e.g., [8]) but were not intended to provide computationally tractable approximations of the marginal likelihood in (1) or the BMSC in (3).

2 Cross-Entropy Laplace Approximation for Bayesian Model Selection

Since $p(\mathcal{D}_n|\boldsymbol{\theta}, \mathcal{M}) \equiv exp(-n\tilde{\ell}_n(\boldsymbol{\theta}))$, in an analogous manner let $p(\ddot{\mathcal{D}}_n|\boldsymbol{\theta}, \mathcal{M}) \equiv exp(-n\ell(\boldsymbol{\theta}))$ be chosen to approximate $p(\mathcal{D}_n|\boldsymbol{\theta}, \mathcal{M})$ for a given sample size n because $\tilde{\ell}_n \rightarrow \ell$ on Θ with probability one under the regularity assumptions $\mathbf{A1} - \mathbf{A5}$ which are provided in the final section of this paper.

Note that when $p(\ddot{\mathcal{D}}_n|\boldsymbol{\theta}, \mathcal{M})$ is substituted for $p(\mathcal{D}_n|\boldsymbol{\theta}, \mathcal{M})$ in (1) to obtain revised formulas for (2) and (3), then (3) may be interpreted as the average amount of ”surprise” or average information received when using model $\mathcal{M}$ to model n samples from the data generating process p_o.

Theorem 2.1 (GBIC Cross-Entropy Approximation)**.** *Assume Assumptions* $\mathbf{A1} - \mathbf{A6}$ *hold. Let the model prior probability density* $p_{\boldsymbol{\theta}} : \Theta \rightarrow [0, \infty)$ *be a continous function on* Θ *such that for all* $\boldsymbol{\theta} \in \Theta$*:* $p_\theta(\boldsymbol{\theta}) > 0$*. Let* $p(\ddot{\mathcal{D}}_n|\mathcal{M}) \equiv exp(-n\ell(\boldsymbol{\theta}))$ *where* $\ell(\boldsymbol{\theta}) \equiv -\int p_o(\mathbf{x}) \log p(\mathbf{x}|\boldsymbol{\theta}) d\nu(\mathbf{x}) < \infty$*. Assume there exists a number* n_0 *such that for all* $n \geq n_0$*:* $p(\ddot{\mathcal{D}}_n|\mathcal{M}) < \infty$*. Then as* $n \rightarrow \infty$*,*

$$-(1/n) \log p(\ddot{\mathcal{D}}_n|\mathcal{M}) = E\{\tilde{\ell}_n(\hat{\boldsymbol{\theta}}_n)\} + (1/(2n)) TRACE \left[(\hat{\mathbf{A}}_n)^{-1} \hat{\mathbf{B}}_n\right] -$$

$$\frac{\log p_\theta(\hat{\boldsymbol{\theta}}_n|\mathcal{M})}{n} + \frac{q}{2n} \log\left(\frac{n}{2\pi}\right) + \frac{\log(\det(\hat{\mathbf{A}}_n))}{2n} + o_p\left(\frac{1}{n}\right). \quad (4)$$

Proof. First, use the Multidimensional Laplace Approximation Theorem ([2], pp. 86-88) leaving $\ell(\boldsymbol{\theta}^*)$, $\mathbf{A}^*$, $\mathbf{B}^*$, and $p_\theta(\boldsymbol{\theta}^*)$ to be estimated. The estimators $\hat{\mathbf{A}}_n = \mathbf{A}^* + o_p(1)$, $\hat{\mathbf{B}}_n = \mathbf{B}^* + o_p(1)$, and $p_\theta(\hat{\boldsymbol{\theta}}_n) = p_\theta(\boldsymbol{\theta}^*) + o_p(1)$ can be substituted to estimate $\mathbf{A}^*$, $\mathbf{B}^*$, and $p_\theta(\boldsymbol{\theta}^*)$ respectively because in conjunction with the existing assumptions the resulting approximation error associated with these substitutions in (4) is $o_p(1/n)$. Second, Proposition P2 of Linhart and Volkers (1984)(see [9]) shows that

$$\ell(\boldsymbol{\theta}^*) = E\{\tilde{\ell}_n(\hat{\boldsymbol{\theta}}_n)\} + (1/(2n)) TRACE \left[(\hat{\mathbf{A}}_n)^{-1} \hat{\mathbf{B}}_n\right] + o_p(1/n). \quad (5)$$

Thus, Equation (5) *must be used* rather than $\ell(\boldsymbol{\theta}^*) = E\{\tilde{\ell}_n(\hat{\boldsymbol{\theta}}_n)\} + O_p(1/n)$ to estimate $\ell(\boldsymbol{\theta}^*)$ to ensure the approximation error in (4) is $o_p(1/n)$. □

This theorem yields a new alternative high-order Laplace Approximation for (1) and (3).

Definition (GBIC Cross-Entropy Approximation (GBIC$_X$))**.** The *Generalized Bayesian Information Criterion Cross Entropy Approximation* is defined as:

$$\text{GBIC}_X = 2n\tilde{\ell}_n(\hat{\boldsymbol{\theta}}_n) + TRACE[\hat{\mathbf{A}}^{-1}\hat{\mathbf{B}}] - 2\log\left[p_\theta(\hat{\boldsymbol{\theta}}_n|\mathcal{M})\right] + q\log\left(\frac{n}{2\pi}\right) + \log(\det(\hat{\mathbf{A}}_n)).$$

3 Simulation Studies

The performance of seven distinct model selection criteria (AIC, GAIC, BIC, GBIC, GBIC$_P$, GBIC$_L$, and GBIC$_X$) and the Negative Normalized Log-Likelihood (NNL) (i.e., $\hat{\ell}_n$) were compared in a series of simulation studies. Note that NNL, AIC, and GAIC are model selection criteria designed to estimate the expected value of NNL, while BIC, GBIC, GBIC$_P$, GBIC$_L$, and GBIC$_X$ are model selection criteria designed to approximate Bayesian Model Selection using the integral in (1). Also note the derivation of AIC assumes the model is correctly specified, while the derivations of the other model selection criteria (i.e., GAIC, BIC, GBIC, GBIC$_P$, GBIC$_L$, and GBIC$_X$) do not assume correct specification.

The performance of the seven model selection criteria was evaluated as follows. A logistic regression model was fitted to a data set of n records. Next, the covariate patterns from the data set of n records were sampled with replacement m times to generate m new data sets. The response variable for each record in each of the m new simulated data sets was then chosen by using the fitted logistic regression model to predict the response variable value probability given each covariate pattern and then randomly setting the response variable value equal to one with the probability specified by the fitted logistic regression model.

The seven model selection criteria were then used to compare the fit of each of the m simulated data sets to the original logistic regression model which was used to generate the response variable values as well as three *decoy models* so that the percentage of correct model selections could be explicitly computed. Since (e.g., [8]) BIC-family criteria (i.e., BIC, BIC, GBIC, GBIC$_P$, GBIC$_L$, and GBIC$_X$) are biased to find models with fewer parameters than AIC-family criteria (i.e., AIC and GAIC), decoy models for the BIC-biased simulation omitted highly predictive covariates to avoid acceptance of decoy models with missing covariates. In contrast, decoy models for the AIC-biased simulation omitted less critical covariates to encourage BIC-family criteria to accept decoy models with missing covariates.

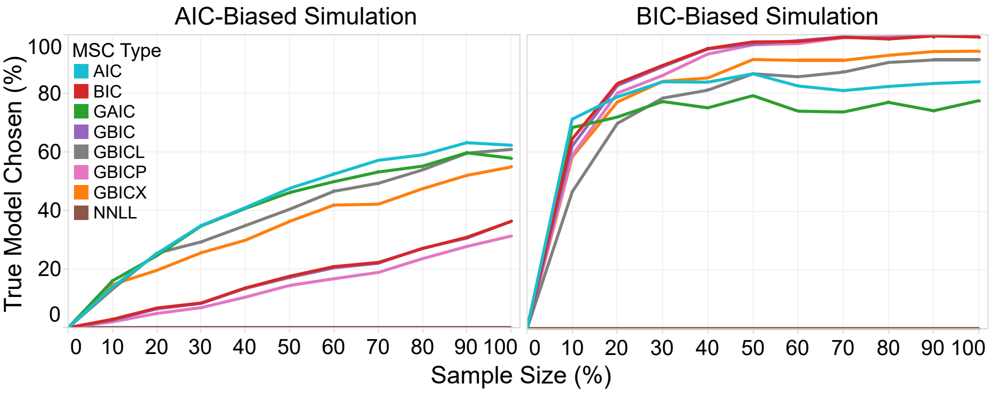

Figure 1: **Percentage of times true model selected as a function of sample size.** The new model selection criterion GBIC$_X$ showed performance which was superior to BIC-family BMSC for AIC-biased simulation studies and showed performance which was superior to the AIC-family model selection criteria for BIC-biased simulation studies.

In the simulations reported here, seven data sets (”car”, n=1728; ”ctg’, n=2126; ”gamma”, n=19020; ”liver”, n= 583; ”news”, n=15000; ”wine”, n=1599; ”white wine”, n=4898) were downloaded from the UCI data repository and prepared for logistic regression modeling by rescaling numerical predictors and removing redundant predictors. Figure (1) shows the average simulation results across the seven data sets which were also observed for the individual seven data sets as well. The empirical results are consistent with the simulation study design. The AIC-family (MSC) exhibited superior performance for the AIC-biased simulation study, while the BIC-family exhibited superior performance for the BIC-biased simulation study. The new model selection criterion, GBIC$_X$ and the version of the classic Laplace approximation model selection criterion GBIC$_L$ exhibited an intermediate robust level of performance for both the AIC-biased and BIC-biased simulation studies. That is, GBIC$_L$ and GBIC$_X$ showed superior performance to BIC, GIBC$_P$, and GBIC$_X$ for the AIC-biased study and also showed superior performance to AIC and $GAIC$ selection criteria for the BIC-biased simulation study. These findings suggest that GBIC$_X$ may be especially useful in situations where a more robust BMSC approximation is desirable.

4 Regularity Assumptions

Assumption A1. The observed data $\mathcal{D}_n \equiv [\mathbf{x}_1, \ldots, \mathbf{x}_n]$ is a realization of sequence of n i.i.d. d-dimensional random vectors with common Radon-Nikodym density p_o with respect to measure ν.

Assumption A2. The learning machine’s environmental model $\mathcal{M} \equiv \{p(\cdot|\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$ is a set of density functions defined with respect to ν in **A1** and closed and bounded parameter space $\Theta \subseteq \mathcal{R}^q$.

Assumption A3. Let Ω be a finite partition of $\mathcal{R}^d$. Let $c : \Omega \times \Theta \rightarrow \mathcal{R}$ be a piecewise continuous function on Ω in its first argument and assume c is twice continuously differentiable in its second argument. The quantity $c(\mathbf{x}, \boldsymbol{\theta})$ is the loss incurred by the learning machine for event $\mathbf{x}$ in its environment and with its parameter values equal to $\boldsymbol{\theta}$.

Assumption A4. The functions c, $\nabla c(\tilde{\mathbf{x}}, \boldsymbol{\theta})$, $\nabla c(\tilde{\mathbf{x}}, \boldsymbol{\theta})[\nabla c(\tilde{\mathbf{x}}, \boldsymbol{\theta})]^T$, and $\nabla^2 c(\tilde{\mathbf{x}}, \boldsymbol{\theta})$ are dominated by integrable functions on Θ with respect to p_o. Sufficient conditions for **A4** to hold are that: (i) the data generating process is a sequence of bounded random vectors (e.g., discrete finite-valued random vectors), and (ii) assumptions $\mathbf{A1} - \mathbf{A3}$ hold.

Assumption A5. There exists a parameter vector $\boldsymbol{\theta}^*$, which is a unique global minimizer of $\ell(\cdot) \equiv E\{c(\tilde{\mathbf{x}}, \cdot)\}$ in the interior of Θ. Note that this does not rule out learning machines with multiple strict local minimizers.

Assumption A6. Let $\mathbf{A} \equiv \nabla^2 \ell$. Assume $\mathbf{A}^* \equiv \mathbf{A}(\boldsymbol{\theta}^*)$ is positive definite. Let $\mathbf{B} \equiv E\left\{\nabla c(\tilde{\mathbf{x}}_i, \cdot)[\nabla c(\tilde{\mathbf{x}}_i, \cdot)]^T\right\}$. The empirical risk estimator $\hat{\boldsymbol{\theta}}_n$ is the unique global minimizer of $\tilde{\ell}_n(\cdot) \equiv (1/n) \sum_{i=1}^{n} c(\tilde{\mathbf{x}}_i, \cdot)$.

References

[1] L. Wasserman. Bayesian model selection and model averaging. *Journal of Mathematical Psychology*, 44:92–107, 2000.

[2] M. Evans and T. Swartz. *Approximating Integrals via Monte Carlo and Deterministic Methods*. Oxford University, Oxford, 2005.

[3] G. Schwarz. Estimating dimension of a model. *Annals of Statistics*, 6(2):461–464, 1978. Times Cited: 7569.

[4] Jinchi Lv and Jun S. Liu. Model selection principles in misspecified models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):141–167, 2014.

[5] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19:716–723, 1974.

[6] Hamparsum Bozdogan. Model selection and akaike's information criterion (aic): The general theory and its analytical extensions. *Psychometrika*, 52(3):345–370, 1987.

[7] K. Takeuchi. Distribution of information statistics and a criterion of model fitting for adequacy of models. *Mathematical Sciences*, 153:12–18, 1976.

[8] Gerda Claeskens and Nils Lid Hjort. *Model selection and model averaging*. Cambridge series in statistical and probabilistic mathematics. Cambridge University Press, Cambridge ; New York, 2008.

[9] H. Linhart and P. Volkers. Asymptotic criteria for model selection. *Operations-Research-Spektrum*, 6(3):161–165, 1984.