A Laplace Approximation for Approximate Bayesian Model Selection

Richard M. Golden (golden@utdallas.edu), Shaurabh Nandy, Vishal Patel, and Pratibha Viraktamath

UT DALLAS The University of Texas at Dallas

NIPS Neural Information Processing Systems

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 alternative approach. In this paper, a computationally intractable multidimensional BMSC integral is approximated using a Laplace approximation 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.

Theorem 2.1 (GBIC Cross-Entropy Approximation). *Assume Assumptions* $\mathbf{A1} - \mathbf{A6}$ *hold. Let the model prior probability density* $p_{\boldsymbol\theta} : \Theta \to [0,\infty)$ *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* $n \to \infty$,

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

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

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 $o_p(1/n)$. □

Theory

Data Set : $\mathcal{D}_n \equiv [\mathbf{x}_1,...,\mathbf{x}_n]$ be a realization of an *i.i.d.* sequence $\tilde{\mathcal{D}}_n \equiv [\tilde{\mathbf{x}}_1,...,\tilde{\mathbf{x}}_n]$ with common density $p_o(\mathbf{x})$.

Probability Model : $\mathcal{M} \equiv \{p(\mathbf{x}|\boldsymbol\theta,\mathcal{M}) : \boldsymbol\theta \in \Theta_{\mathcal{M}} \subset \mathcal{R}^q\}$

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

$\tilde{l}_n(\boldsymbol\theta;\mathcal{M}) \equiv -(1/n)\log p(\mathcal{D}_n|\boldsymbol\theta,\mathcal{M})$, $\hat{\boldsymbol\theta}_n \equiv \arg\min_{\boldsymbol\theta\in\Theta_{\mathcal{M}}} \tilde{l}_n(\boldsymbol\theta;\mathcal{M})$

$l(\boldsymbol\theta;\mathcal{M}) \equiv E\{\tilde{l}_n(\boldsymbol\theta;\mathcal{M})\}$, $\boldsymbol\theta^* \equiv \arg\min_{\boldsymbol\theta\in\Theta_{\mathcal{M}}} l(\boldsymbol\theta;\mathcal{M})$

$\tilde{\mathbf{A}}_n \equiv \nabla^2 \tilde{l}_n(\hat{\boldsymbol\theta}_n;\mathcal{M})$, $\tilde{\mathbf{B}}_n \equiv (1/n)\sum_{i=1}^{n} \nabla \log p(\mathbf{x}_i|\hat{\boldsymbol\theta}_n,\mathcal{M})(\nabla \log p(\mathbf{x}_i|\hat{\boldsymbol\theta}_n,\mathcal{M}))^T$

$p(\mathcal{D}_n|\boldsymbol\theta,\mathcal{M}) = \exp(-n\tilde{l}_n(\boldsymbol\theta;\mathcal{M}))$, $p(\ddot{\mathcal{D}}_n|\boldsymbol\theta,\mathcal{M}) = \exp(-nl(\boldsymbol\theta;\mathcal{M}))$

Marginal Likelihood for $\mathcal{M}$: $p(\mathcal{D}_n|\mathcal{M}) \equiv \int p(\mathcal{D}_n|\boldsymbol\theta,\mathcal{M})\, p_\theta(\boldsymbol\theta|\mathcal{M}) d\boldsymbol\theta$ with prior $p_\theta(\boldsymbol\theta|\mathcal{M})$

AIC $\equiv 2n\tilde{l}_n(\hat{\boldsymbol\theta}_n;\mathcal{M}) + 2q = -2nE\{\log p(\tilde{\mathcal{D}}_n|\boldsymbol\theta^*,\mathcal{M})\} + o_p(1)$ *only if* $p_o \in \mathcal{M}$ (Akaike, 1974)

GAIC $\equiv 2n\tilde{l}_n(\hat{\boldsymbol\theta}_n;\mathcal{M}) + 2TRACE((\tilde{\mathbf{A}}_n)^{-1}\tilde{\mathbf{B}}_n) = -2nE\{\log p(\tilde{\mathcal{D}}_n|\boldsymbol\theta^*,\mathcal{M})\} + o_p(1)$ (Takeuchi, 1976)

BIC $= 2n\tilde{l}_n(\hat{\boldsymbol\theta}_n;\mathcal{M}) + q\log(n) = -2\log p(\mathcal{D}_n|\mathcal{M}) + O_p(1)$ (Schwarz,1978)

$\mathbf{GBIC}_L = 2n\tilde{l}_n(\hat{\boldsymbol\theta}_n;\mathcal{M}) - 2\log p_\theta(\hat{\boldsymbol\theta}_n|\mathcal{M}) + q\log\left(\frac{n}{2\pi}\right) + \log\det\tilde{\mathbf{A}}_n = -2\log p(\mathcal{D}_n|\mathcal{M}) + o_p(1)$ (e.g., Wasserman, 2000)

$\mathbf{GBIC} = 2n\tilde{l}_n(\hat{\boldsymbol\theta}_n;\mathcal{M}) - 2\log p_\theta(\hat{\boldsymbol\theta}_n|\mathcal{M}) + q\log\left(\frac{n}{2\pi}\right) - \log\det((\tilde{\mathbf{A}}_n)^{-1}\tilde{\mathbf{B}}_n) = -2\log p(\mathcal{D}_n|\mathcal{M}) + o_p(1)$ (Lv and Liu, 2014)

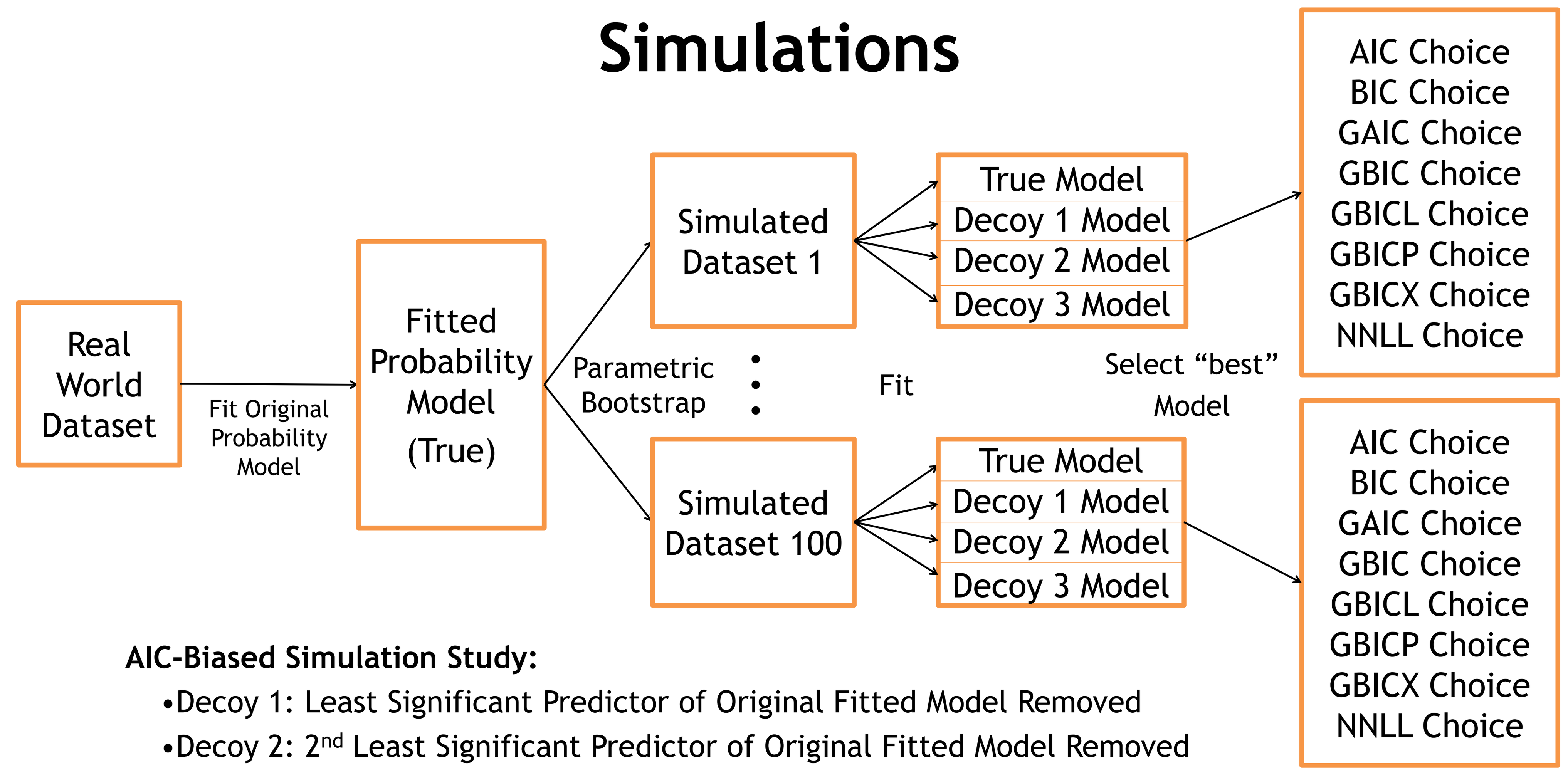
$\mathbf{GBIC}_P = 2n\tilde{l}_n(\hat{\boldsymbol\theta}_n;\mathcal{M}) - 2\log p_\theta(\hat{\boldsymbol\theta}_n|\mathcal{M}) + q\log\left(\frac{n}{2\pi}\right) - \log\det((\tilde{\mathbf{A}}_n)^{-1}\tilde{\mathbf{B}}_n) + TRACE((\tilde{\mathbf{A}}_n)^{-1}\tilde{\mathbf{B}}_n)$

$= -2\log p(\mathcal{D}_n|\mathcal{M}) + o_p(1)$ (Lv and Liu, 2014)

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

$= -2\log p(\ddot{\mathcal{D}}_n|\mathcal{M}) + o_p(1)$ (**New Result**!)

Simulations



AIC-Biased Simulation Study:

- Decoy 1: Least Significant Predictor of Original Fitted Model Removed
- Decoy 2: 2nd Least Significant Predictor of Original Fitted Model Removed
- Decoy 3: Include Product (Interaction) of Least Significant and 2nd Least Significant Predictor

BIC-Biased Simulation Study:

- Decoy 1: Most Significant Predictor of Original Fitted Model Removed
- Decoy 2: 2nd Most Significant Predictor of Original Fitted Model Removed
- Decoy 3: Include Product (Interaction) of Least Significant and 2nd Least Significant Predictor

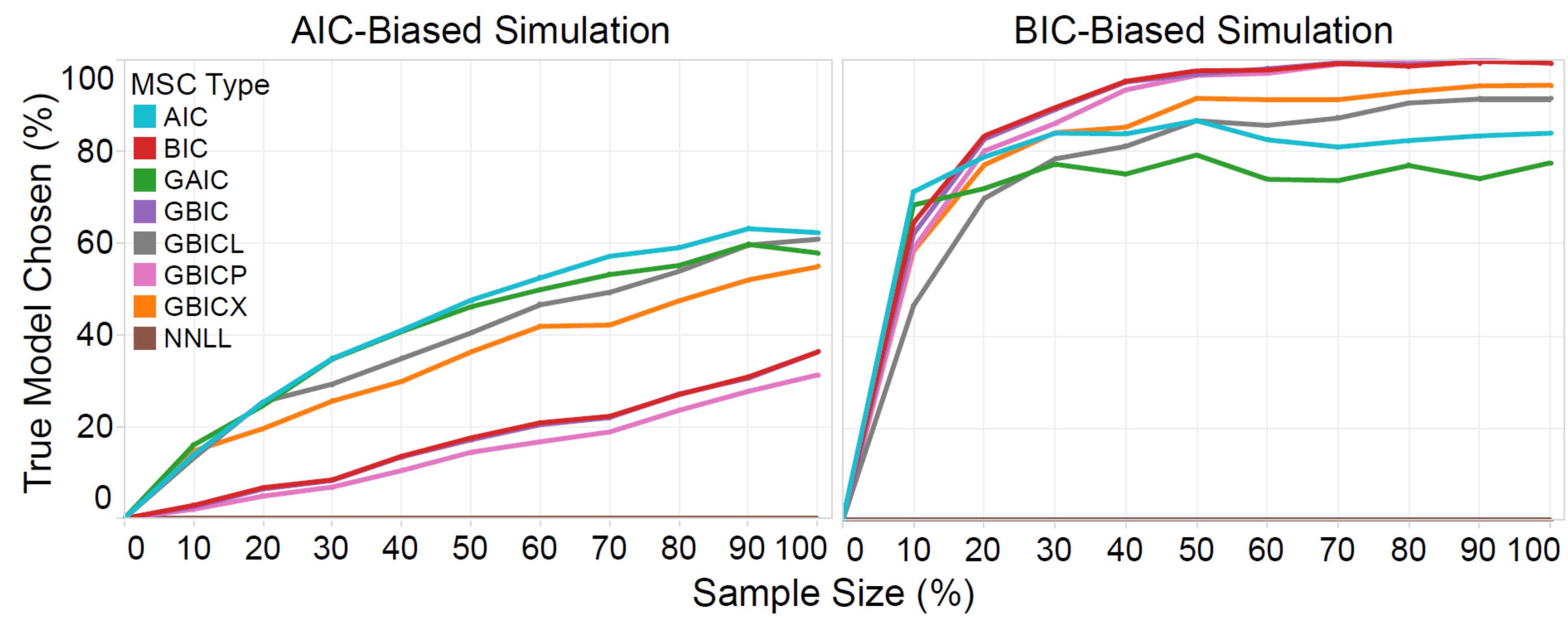


Figure 1: **Percentage of times true model selected as a function of sample size.** The new model selection criterion GBIC_X showed performance which was superior to BIC-family BMSC for AIC-biased simulation studies and showed performance which was superior to the AIC-family model selection criteria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