Available online at www.sciencedirect.com
SCIENCE DIRECT®
ELSEVIER
Computational Statistics & Data Analysis 48 (2005) 221–234
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Assessment of two approximation methods for computing posterior model probabilities

Edward L. Boone*, Keying Ye, Eric P. Smith

Department of Mathematics and Statistics, UNC-Wilmington, 601 S. College Rd., Wilmington 28403, USA

Received 28 August 2003; received in revised form 20 January 2004; accepted 21 January 2004

Abstract

Model selection is an important problem in statistical applications. Bayesian model averaging provides an alternative to classical model selection procedures and allows researchers to consider several models from which to draw inferences. In the multiple linear regression case, it is difficult to compute exact posterior model probabilities required for Bayesian model averaging. To reduce the computational burden the Laplace approximation and an approximation based on the Bayesian information criterion (BIC) have been proposed. The BIC approximation is the easiest to calculate and is being used widely in application. In this paper we conduct a simulation study to determine which approximation performs better. We give an example of where the methods differ, study the performance of these methods on randomly generated models and explore some of the features of the approximations. Our simulation study suggests that the Laplace approximation performs better on average than the BIC approximation.

Keywords: Laplace approximation; Schwarz Bayesian information criterion; Bayesian model averaging; Model selection; Variable assessment

1. Introduction

Model building in multiple linear regression (MLR) often requires assessing which subset of variables create the "best" model. Methods such as Stepwise, Forward, Backward selection, PRESS, Mallow's C_p, Akaike's Criterion, Schwartz's Bayesian Information Criterion have been developed to address this issue. Once a "best" model is

* Corresponding author. Tel.: +1-910-962-3298; fax: +1-910-962-7107.
E-mail address: eboone@vt.edu (E.L. Boone).

0167-9473/$ - see front matter © 2004 Published by Elsevier B.V.
doi:10.1016/j.csda.2004.01.005

selected all inferences are based on this model as if it were the true model, no uncertainty is associated with the modelling process. In the recent literature efforts have been made to incorporate this model uncertainty into the analysis using posterior model probabilities (see Kass and Raftery 1995; Madigan and York, 1995; Raftery, 1996; Raftery et al., 1997; Hoeting et al., 1999).

Posterior model probabilities have many uses in modelling. They can be used to select the highest probability model, to aid variable assessment and for prediction. The main justification for using any of the aforementioned methods is to formally incorporate model uncertainty into the analysis. By incorporating this uncertainty we arrive at inferences that more accurately reflect all the uncertainties associated with the analysis. This type of analysis has been growing in popularity in application. For example, Murphy and Wang (2001) considered an infant survival application. Viallefont et al. (2001) employed these methods on case–control studies. In addition, Clyde (2000), Lamon and Clyde (2000), Noble (2000) and Lipkovich (2002) explored ecological data sets with these methods.

In this paper we examine the performance of two methods for approximating posterior model probabilities, the Laplace approximation and the Bayesian information criterion (BIC) approximation method. We focus on the *multiple linear regression* setting to evaluate the performance. Via simulation and example we show that the Laplace approximation is more accurate than BIC with moderate to large sample sizes. We then employ the Laplace approximation, BIC method and the Exact method to an ecological data set collected by the State of Ohio Environmental Protection Agency. We use this case study to illustrate the importance of accurate posterior model probability approximations. This case study suggests the relative the poor performance of the BIC method in practice.

Computing the exact posterior model probabilities is computationally intensive. The calculation involves the inverse and determinant of an $n \times n$ matrix where n is the sample size. This calculation needs to be repeated for each model. In the situation where we have k candidate variables, there are 2^k models to consider. Hence, the inverse and determinant would need to be calculated 2^k times. As n grows large the computation time becomes infeasible. Hence approximations are needed to reduce the computation time.

From experience, we have observed disagreement between the BIC and the Laplace method for calculating posterior model probabilities. Weakliem (1999) noted that the BIC approximation to Bayes factors is too conservative. Raftery (1999) stated that the BIC approximation method is a "crude" approximation, however, BIC values are readily available on standard computer output. Noble (2000) observed poor performance of the BIC method for small sample sizes and suggested corrections to improve performance. Noble's adjustment reduced the penalty associated with BIC. Noble (2000) did not directly compare the BIC method or the Laplace method to the exact probabilities. Both the BIC method and the Laplace method are asymptotically accurate methods for determining Bayes factors, which can be used to determine the posterior model probabilities (Tierney and Kadane, 1986; Raftery, 1996; Kass and Wasserman, 1995). Chickering and Heckerman (1997) consider the performance of the BIC and Laplace methods for finite mixture models and find the BIC method does not perform well

in that case. Our focus is on the multiple linear regression setting. Since these are asymptotic approximations, we need to understand how these methods perform in small to moderate sample sizes in order to know when to apply each method.

In this paper, we present an extensive simulation study of the accuracy of using the BIC and Laplace approximation in the context of Bayesian model averaging (BMA). In the remainder of this section we discuss BMA, the BIC and Laplace approximations in this context. The next section contains an example where the BIC and Laplace methods lead to different conclusions when used for single model selection via the highest probability model and when used for variable assessment. In Section 3, we consider the performance for randomly generated models. We conduct a study of the accuracies in the random case by varying the number of "significant" variables in the model in Section 4.

All algorithms were coded in Microsoft Visual Basic for Applications. We use the RanDev.dll library for all psuedo-random number generation. This requires the programs to run in a PC environment. All programs are available from the authors.

1.1. Bayesian model averaging

Suppose we have k candidate predictor variables $X_1, X_2, \ldots, X_k$ and a single response Y. In this situation we have 2^k first-order models which can be formed from these predictors. Let $\mathscr{M}$ be the set of all possible models and let M_i denote the ith model in the set $\mathscr{M}$. The cardinality or size of $\mathscr{M}$ is denoted by $|\mathscr{M}|$.

Once we collect data $\mathbf{D}$, we can determine for model $M_i \in \mathscr{M}$, $P(M_i|\mathbf{D})$ the posterior probability of M_i:

$$P(M_i|\mathbf{D}) = \frac{P(M_i)P(\mathbf{D}|M_i)}{\sum_{M_j \in \mathscr{M}} P(M_j)P(\mathbf{D}|M_j)}, \tag{1}$$

where $P(M_i)$ is the prior probability of model M_i and

$$P(\mathbf{D}|M_i) = \int L(\mathbf{D}|\theta_i, M_i)P(\theta|M_i)\,\mathrm{d}\theta \tag{2}$$

with $L(\mathbf{D}|\theta_i, M_i)$ being the likelihood of the data given the parameter vector θ_i for model M_i and $P(\theta|M_i)$ being the prior probability density for θ_i given model M_i. The calculation of the exact quantities in Eq. (1) is computationally intensive.

For large model spaces, $P(M_i|\mathbf{D})$ can be directly estimated by the Markov Chain Monte Carlo methods used by Madigan and York (1995), Raftery et al. (1997), Noble (2000) and Lipkovich (2002). To accomplish this, a Markov Chain is created where each state is a model M. Transitions from model M_i to model M_j are governed by the following acceptance probability α:

$$\alpha = \min\left(1, \frac{P(M_j|\mathbf{D})}{P(M_i|\mathbf{D})}\right).$$

This method depends on the accuracy of $P(M_i|\mathbf{D})$. Noble (2000) and Lipkovich (2002) use the BIC approximation to determine $P(M_i|\mathbf{D})$.

The accuracy of $P(M_i|\mathbf{D})$, for a given data set $\mathbf{D}$, is crucial to ensure the posterior inferences are correct. In the model selection context, the model M_i with the highest values of $P(M_i|D)$ is selected to be the best model. In the model averaging context, for a quantity Δ, the posterior model probabilities are used in the law of total probability formula

$$P(\Delta|\mathbf{D}) = \sum_{i=1}^{|\mathscr{M}|} P(\Delta|M_i, \mathbf{D})P(M_i|\mathbf{D}). \tag{3}$$

In the regression setting our parameter vector is $\boldsymbol{\theta}_i = (\boldsymbol{\beta}_i, \sigma^2)$, where $\boldsymbol{\beta}_i$ is the regression coefficient vector with $\beta_{ij} = 0$ if $X_j \notin M_i$. The prior probability density for $\boldsymbol{\theta}_i$ needs to be proper (i.e. $\int p(\theta_i, M_i)\,\mathrm{d}\theta_i = 1$) in order for Eq. (1) to exist.

1.2. Calculation methods

For linear models $Y = \mathbf{X}\boldsymbol{\beta}$ Eq. (2) can only be determined analytically for a few special prior distributions. If we cannot obtain (2) analytically, approximations are needed or numerical integration can be employed. In the special case of a normal error regression model, the probabilities expressed in (2) can be analytically determined when a normal prior is used for the β's and a gamma prior for $1/\sigma^2$ (see Raiffa and Schlaifer, 1961). Suppose for each model M_i the normal-gamma prior is used, i.e. $\boldsymbol{\beta}_i$ is normally distributed with some mean μ_i and variance V_i and $\sigma_i^2 \sim \lambda v/\chi_v^2$ where λ is the mean and v represents the degrees of freedom. In this situation we can determine (2) by the following:

$$P(\mathbf{D}|\mu_i, \mathbf{V}_i, v, \mathbf{X}_i, M_i) = \frac{\Gamma((v+n)/2)(v\lambda)^{v/2}}{\pi^{n/2}\Gamma(v/2)|\mathbf{I} + \mathbf{X}_i\mathbf{V}_i\mathbf{X}_i'|^{1/2}} \times [\lambda v + (\mathbf{Y} - \mathbf{X}_i\mu_i)'(\mathbf{I} + \mathbf{X}_i\mathbf{V}_i\mathbf{X}_i')^{-1}(\mathbf{Y} - \mathbf{X}_i\mu_i)^{-(v+n)/2}].$$

where $\mathbf{X}_i$ is the corresponding design matrix for model M_i, and $\Gamma(\cdot)$ is the gamma function. For moderate to large model spaces or large sample sizes this method is computationally intensive since the determinant and the inverse of the $n \times n$ matrix $\mathbf{I} + \mathbf{X}_i\mathbf{V}_i\mathbf{X}_i'$ must be computed for each model. This limits its usability in large sample size problems.

Tierney and Kadane (1986) proposed the idea of using the Laplace approximation to evaluate (2). They showed that this approximation is of order $\mathrm{O}(1/\sqrt{n})$. The Laplace approximation for a unimodal function $p(\theta)$ is given by

$$\int p(\theta)\,\mathrm{d}\theta \approx p(\hat{\theta})(2\pi)^{d/2}|I(\hat{\theta})^{-1}|^{1/2}, \tag{4}$$

where $\hat{\theta}$ is the mode of the probability function $p(\theta)$, d is the dimensionality of $p(\theta)$ and $I(\hat{\theta})$ is the observed information matrix evaluated at $\hat{\theta}$. Standard statistical software cannot readily calculate the posterior mode, nor the information matrix of a posterior distribution. However, the mode $\hat{\theta}$ can easily be calculated by using methods such as Newton–Rhaphson, Fisher's scoring or steepest ascent or any combinations of these.

Raftery (1996), Kass and Raftery (1995) and Kass and Wasserman (1995) suggest approximating (2) using a function of the BIC developed by Schwartz (1978)

$$P(\mathbf{D}|M_i) \approx \mathrm{e}^{1/2BIC_i},$$

where $BIC_i = 2\{\ln p(\mathbf{D}|\hat{\theta}_k, M_k) - d\ln(n)$ is the Bayesian information criterion for model M_k and d is the dimension of the model (Schwartz, 1978). Raftery (1996) and Kass and Wasserman (1995) showed that this approximation is asymptotically accurate for computing a Bayes factor for nested hypotheses and is of order O(1) under the unit-information prior distribution on $\boldsymbol{\beta}$. A "unit information prior" is where the prior distribution for $\boldsymbol{\beta}$ contains the amount of information on $\boldsymbol{\beta}$ as is available in one observation. In the normal case, the unit information prior for testing a nested null hypothesis, H_0: $\phi = \phi_0$, we have $\phi \sim N(\phi_0, \Sigma_\phi)$ where $|\Sigma_\phi|^{-1} = |I_{\phi\phi}(\boldsymbol{\beta}, \phi_0)|$, where $I_{\phi\phi}(\boldsymbol{\beta}, \phi_0)$ denotes the sub-matrix of the Fisher information matrix corresponding to ϕ in the restricted likelihood. The popularity of this approximation is due to the fact that it can be calculated from the BIC values conveniently available from standard statistical software output. Extensions of this approximation for multivariate analysis have been studied by Noble (2000) for principal components analysis. Lipkovich (2002) used this method for canonical correspondence analysis and cluster analysis.

1.3. Comparison criteria

As mentioned earlier, the BIC and Laplace methods are asymptotically valid for determining Bayes factors. However, we usually do not know the sample size required for an accurate approximation, although both Laplace and BIC methods are correct. We are interested in which criterion yields better quality results, under the same conditions. To measure the accuracy of the approximations three measures are used: weighted L_1 and L_2 distances and the Hellinger distance. The L_2 distance of the probability function P_1 relative to the probability function P_2 is given by

$$L_2(P_1, P_2) = \left\{ \sum_{M_i \in \mathscr{M}} [P_1(M_i|\mathbf{D}) - P_2(M_i|\mathbf{D})]^2 P_2(M_i|\mathbf{D}) \right\}^{1/2}.$$

The L_1 distance is given by

$$L_1(P_1, P_2) = \sum_{M_i \in \mathscr{M}} |P_1(M_i|\mathbf{D}) - P_2(M_i|\mathbf{D})| P_2(M_i|\mathbf{D}).$$

Finally, the discrete Hellinger distance is given by

$$d_H(P_1, P_2) = \left\{ 2 - 2 \sum_{M_i \in \mathscr{M}} [P_1(M_i|\mathbf{D}) P_2(M_i|\mathbf{D})]^{1/2} \right\}^{1/2}.$$

The weighted L_1 and L_2 measures, give more importance to and have greater accuracy for models of high probability with respect to P_2. We also chose the Hellinger distance to consider a non-weighted distance measure. In many cases $P(M_i|\mathbf{D}) \approx 0$, thus the discrete analog of the Kullback–Liebler distance is not appropriate. With all of these

distance measures values closer to zero correspond to a smaller distance between distributions. The L_1 and L_2 distances are bounded between 0 and 1 and the Hellinger distance is bounded between 0 and 2. To evaluate the methods, we deem distances less than 0.05 to be acceptable for L_1 and L_2 and less than 0.2 for the Hellinger distance. These values correspond to 95% of the total probability being assigned correctly.

2. Example

To illustrate the importance of the accuracy of both methods, we employ each method on an environmental data set collected in Ohio. In this section we examine the highest probability models, and variable assessment of each of the methods on a biological data set provided by the Ohio Environmental Protection Agency (EPA). The Ohio EPA is interested in the health of the fish living in the streams and rivers of Ohio and how this is affected by environmental stress. It is especially important how habitat and chemical variables affect the health of these fish. The response measure is the index of biotic integrity (IBI) which is a composite measure of the health of the fish community. The predictors are quantitative habitat evaluation index (QHEI) which measures the quality of the habitat, dissolved oxygen, hardness, pH and total alkalinity. It is important to know which of these predictors influences the health of the fish. By understanding what affects the health of the fish, regulators can set guidelines to improve the health of the fish population.

In this analysis we used the normal-gamma prior distribution for the parameters with $\beta_i \overset{\text{iid}}{\sim} \text{N}(0, 100)$ and $\sigma^2 \sim Inv - \chi^2(2)$. A diffuse prior is appropriate in this setting since we do not have any prior information about the parameters. Since the sample size is $n = 125$ and we have five variables, the prior distribution should not have much effect on the posterior inferences. A uniform prior distribution was placed on the model space $\mathscr{M}$ to reflect our prior uncertainty about which model is the correct model.

When a researcher is interested in selecting a single model, one can select the model with the highest posterior probability. Table 1 shows the 10 highest probability models using the exact method and the corresponding posterior model probabilities generated by each method. In our example, the exact and Laplace approximation select the same model with only QHEI in the model, the BIC method allowed QHEI, DO and hardness into the highest probability model. This discrepancy among highest probability models illustrates the concern with using the BIC approximation. We also notice that the Laplace approximation places more weight on the model with QHEI only than does the exact method.

One use of BMA is variable assessment. In variable assessment we obtain the posterior probability the coefficient of a variable is different from zero. Mathematically, we have $P(\beta_i \neq 0|\mathbf{D}) = \sum_{M \in \mathscr{M}} I_{X_i}(M) P(M|\mathbf{D})$, where $I_{X_i}(M)$ is an indicator function which takes on value 1 when variable X_i is in model M and 0 otherwise. In this situation, high posterior probabilities correspond to variables being important. General guidelines for interpreting these probabilities are given by Kass and Raftery (1995) which is shown in Table 2. In Table 3 the probabilities are given for our data set.

Table 1
The 10 highest posterior probability models using exact method and the corresponding posterior model probabilities from exact, Laplace and BIC methods

QHEI	DO	Hardness	pH	Alkalinity	Exact	Laplace	BIC
X					**0.492**	**0.608**	0.089
X	X				0.204	0.167	0.156
X	X	X			0.145	0.069	**0.448**
X		X			0.058	0.055	0.048
X				X	0.025	0.026	0.017
X			X		0.022	0.025	0.008
X	X		X		0.011	0.005	0.019
X	X			X	0.009	0.003	0.023
X	X	X	X		0.008	0.001	0.061
X	X	X		X	0.005	0.010	0.059

Table 2
Guidelines for interpreting activation probabilities $P(\beta_i \neq 0|\mathbf{D})$

$p = P(\beta_i \neq 0|\mathbf{D})$	Evidence
$p \leqslant 0.5$	None
$0.5 < p \leqslant 0.75$	Mild
$0.75 < p \leqslant 0.95$	Positive
$0.95 < p \leqslant 0.99$	Strong
$p > 0.99$	Very strong

Table 3
Variable assessment using Exact, Laplace and BIC methods

Variable	Exact	Laplace	BIC
QHEI	0.990	0.978	0.995
DO	0.389	0.265	0.821
Hardness	0.226	0.139	0.679
pH	0.047	0.042	0.106
Alkalinity	0.044	0.039	0.121

Using the guidelines from Table 2, we see that both the exact method and the Laplace approximation place QHEI in the strong or very strong evidence category, and all other variables in the no evidence category. Using the BIC approximation, QHEI shows positive evidence, DO and hardness show mild evidence and all others show no evidence. Selecting QHEI as the only important variable has important management implications and would lead to a focus on improving habitat rather than controlling chemical and fertilizer stress. This example shows the need for the study of the accuracy of the BIC and Laplace approximations.

Table 4
Mean distance between exact posterior distribution and the BIC and Laplace approximations to the posterior distribution using random underlying models

n	L_2 Mean		L_1 Mean		d_H Mean	
	BIC	Laplace	BIC	Laplace	BIC	Laplace
50	0.1323	0.0881	0.1276	0.0846	0.1975	0.1444
	(0.0067)	(0.0060)	(0.0065)	(0.0058)	(0.0102)	(0.0080)
75	0.0972	0.0324	0.0947	0.0317	0.1503	0.0764
	(0.0066)	(0.0041)	(0.0065)	(0.0040)	(0.0084)	(0.0053)
100	0.0874	0.0324	0.0850	0.0317	0.1353	0.0514
	(0.0069)	(0.0030)	(0.0067)	(0.0030)	(0.0087)	(0.0039)
125	0.0845	0.0230	0.0826	0.0224	0.1267	0.0356
	(0.0068)	(0.0024)	(0.0067)	(0.0023)	(0.0088)	(0.0030)
150	0.0665	0.0184	0.0653	0.0181	0.0986	0.0278
	(0.0065)	(0.0021)	(0.0063)	(0.0020)	(0.0080)	(0.0026)
175	0.0630	0.0120	0.0611	0.0117	0.0936	0.0192
	(0.0066)	(0.0015)	(0.0065)	(0.0015)	(0.0085)	(0.0020)
200	0.0587	0.0115	0.0575	0.0112	0.0896	0.0179
	(0.0062)	(0.0015)	(0.0061)	(0.0014)	(0.0080)	(0.0019)

Standard errors are shown in parenthesis below the mean.

3. Random models

To understand how the methods perform in general we conducted the following simulation study. We allowed both $\boldsymbol{\beta}$ and σ^2 to be random. We used the 5 regressor case with $X_{ij} \overset{\text{iid}}{\sim} \text{N}(0,1)$ for all $i = 1,\ldots,5$ and $j = 1,\ldots,n$. We used the following distributions to sample the parameter values: $\boldsymbol{\beta}_i \overset{\text{iid}}{\sim} \text{N}(0,4)$ and $\sigma^2 \sim \chi^2(1)$. To sample the random error ε_j we first sampled σ^2 and then sampled $\varepsilon_j|\sigma^2 \overset{\text{iid}}{\sim} \text{N}(0,\sigma^2)$.

We again used a normal-gamma prior distribution for the regression parameters. For $\boldsymbol{\beta}$ we used $\beta_i \overset{\text{iid}}{\sim} \text{N}(0,10)$, reflecting a priori uncertainty about the regression coefficients. We chose $\sigma^2 \sim Inv-\chi^2(2)$ due to the fact that this distribution has infinite variance. For consistency we employed the uniform prior distribution over the model space.

We considered sample sizes of 50, 75, 100, 125, 150 and 200. Each of these simulations was performed 200 times and the average distance was calculated using L_2, L_1 and d_H. Table 4 shows the results of the simulations comparing the Laplace approximation and the BIC approximations to the exact posterior distribution. We did not consider sample sizes above 200 due to the excessive computational time required.

The results in Table 4 show both the Laplace and BIC methods improve as the sample size increases. However, the Laplace method is much closer to the true value, on average, than the BIC method. Furthermore, for these sample sizes, the Laplace approximation is also closer to the exact model probability distribution for a much lower sample size. Since the Laplace approximation performs better, on average, than does the BIC method we feel the Laplace method is preferred to the BIC method.

Table 5
The mean L_1 distance between the exact posterior distribution and the BIC and Laplace approximations to the posterior distribution

n	df	$\sigma_{\beta_i} = 10$		$\sigma_{\beta_i} = 100$	
		BIC	Laplace	BIC	Laplace
50	1	0.128	0.085	0.190	0.117
		(0.007)	(0.006)	(0.012)	(0.009)
	4	0.113	0.083	0.195	0.116
		(0.007)	(0.006)	(0.012)	(0.008)
100	1	0.085	0.032	0.151	0.041
		(0.007)	(0.003)	(0.012)	(0.004)
	4	0.073	0.028	0.139	0.040
		(0.006)	(0.003)	(0.011)	(0.004)
150	1	0.065	0.018	0.138	0.026
		(0.006)	(0.002)	(0.012)	(0.003)
	4	0.071	0.016	0.127	0.025
		(0.006)	(0.002)	(0.012)	(0.003)

The prior regression parameter standard deviation $\sigma_{\beta_i} = 10, 100$ and regression error degrees of freedom $df = 1, 4$. Standard errors are shown in parentheses below the mean.

To assess the sensitivity of the results we considered the prior regression parameter standard deviation, $\sigma_{\beta_i} = 10, 100$. We also varied the regression error, ε_j, degrees of freedom $df = 1, 4$. Table 5 shows the mean L_1 distance for sample sizes of 50, 100, and 150. The results show that while the simulation is sensitive to the prior distribution and regression error degrees of freedom for small sample sizes, the pattern noted above still persists for all sample sizes and parameter settings. We also performed another study varying the degrees of freedom for the prior distribution of σ^2. The results of this study the results were very similar. This suggests that the results are robust to the prior parameter and simulation parameter settings.

4. Exploration

We also wish to understand how each method performs when the number of "significant" variables varies. For this we use the five variable case as before. To determine whether each variable was "significant" we used the standard t-statistic used for testing regression coefficients. For each of the variables we chose the cut off value of 2.5 since this roughly corresponds to the cut off value for five simultaneous tests using the Bonferroni correction at the $\alpha = 0.05$ level. Hence, if the t-statistic for the variable was larger than 2.5 we deemed the variable as significant. We would expect to see the variability of the distances increase as the number of significant variables in the model increases. This is due to the power of the test at each of the sample size levels and should diminish as the sample size increases. For larger sample sizes we would

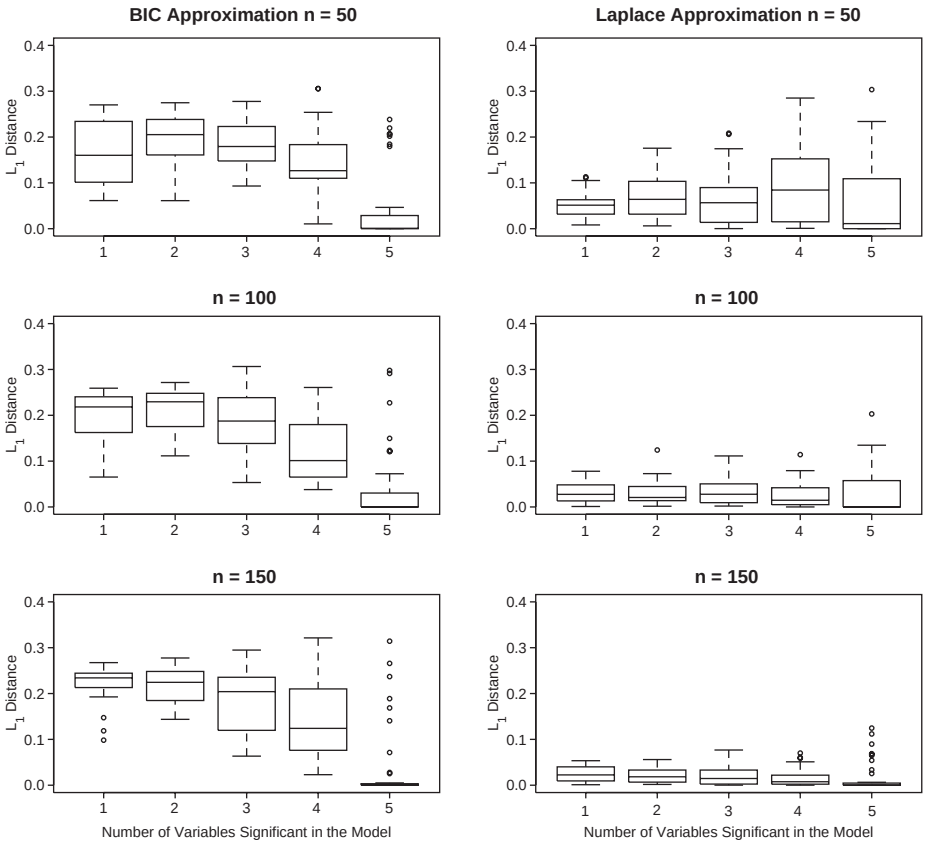

Fig. 1. Box plots of L_1 distances for sample sizes of $n = 50, 100, 150$ for five candidate variables. Distances generated by comparing the BIC approximation to the exact posterior distribution are in the left column and the distances generated by Laplace approximation are in the right column.

expect the distances to be concentrated near 0 with small variability. To study this we obtained 200 simulation samples each for $n = 50$, 100 and 150. The study was controlled to have 40 simulations each with one variable significant, 40 with two variables significant and so on. For each of these models we obtained the L_1, L_2 and Hellinger distances from the exact posterior distribution, using both the BIC approximation and the Laplace approximation.

In Fig. 1 we see the box plots of the L_1 distances for both the BIC and Laplace approximations at the three sample sizes. The plots on the left correspond to the BIC approximation with the Laplace approximation on the right. When examining the Laplace approximation, at $n = 50$ we see the expected pattern, the variability of the distances increased as the number of significant models increased. However, this pattern is not present at $n = 100$ or at $n = 150$. Furthermore, we notice the samples tend to be concentrated at low values. We can also notice that the variability in the distances decreases

as the sample size increases. This same pattern is also exhibited with both the L_2 and Hellinger distances.

By examining the BIC approximation box plots in Fig. 1, we notice that the BIC approximation has large distances from the exact posterior distribution when one, two, three and four variables are significant in the model. Furthermore, the distances for these samples are not concentrated near 0. This pattern persists through all three sample size levels. We also notice it is quite accurate when five variables are significant in the model. This same pattern is also exhibited with both the L_2 and Hellinger distances. Hence, the accuracy of the BIC approximation seems to depend on the number of significant variables in the model. This is an unattractive feature exhibited by the BIC approximation. We would hope that the distances would be concentrated near 0 regardless of the number of significant variables in the model. In addition, we would hope the variation in the distances would decrease with increasing sample size. This pattern was not indicated by the simulations performed in Section 3, due to the fact that as the sample sizes increase, the number of significant variables in the model increased due to increased power. Hence more of the models sampled had large numbers of significant variables in the model.

After noticing this odd pattern produced by the BIC approximation, we wished to understand if this pattern holds for large sample sizes. To study this we focused on the situation where only one variable was significant in the model. We obtained 200 samples for $n = 1000$. The sampling scheme was the similar as used in the previous sections, except we set $\beta_2, \beta_3, \beta_4, \beta_5 = 0.001$ for all of the models. This ensured that only one variable was "significant" in the model. We then computed the L_1, L_2 and Hellinger distances from the Laplace approximation for each of the samples. We used the Laplace approximation as a reference point to avoid the computational issues associated with computing the exact posterior distributions. The Laplace approximation should be a reasonable reference. Using the samples we obtained the L_1 distance mean of 0.1324 with a standard error of 0.0036. The L_2 and Hellinger distances produced similar results. This is an indication that the pattern persists in large sample sizes.

To determine whether our conclusions for other numbers of candidate variables we performed a similar simulation study with three candidate variables. We obtained 200 samples with 66 samples with one "significant" variable, 66 samples with two "significant variables" and 67 with three "significant" variables using the cut off value of 2.4 which roughly corresponds to the Bonferroni correction for three simultaneous tests at the $\alpha = 0.05$ level. We performed these simulations for sample sizes of 30, 75 and 150. We chose 30 as our lower number since it corresponds with the popular rule of thumb that 10 observations are needed for each covariate. Again, we obtained the L_1, L_2 and Hellinger distances of BIC approximation and the Laplace approximation from the exact posterior distribution. Fig. 2 shows the box plots of the L_1 distances for this scenario. The graphs in the left column correspond to BIC approximation and those in the right column to the Laplace approximation. In this we see that the same patterns are exhibited for three variables as well.

To determine if this pattern exists with large sample sizes we obtained the 200 samples for $n = 1000$ and computed the distances between approximation methods. In each of the models sampled $\beta_2, \beta_3 = 0.001$ to ensure that only one variable was

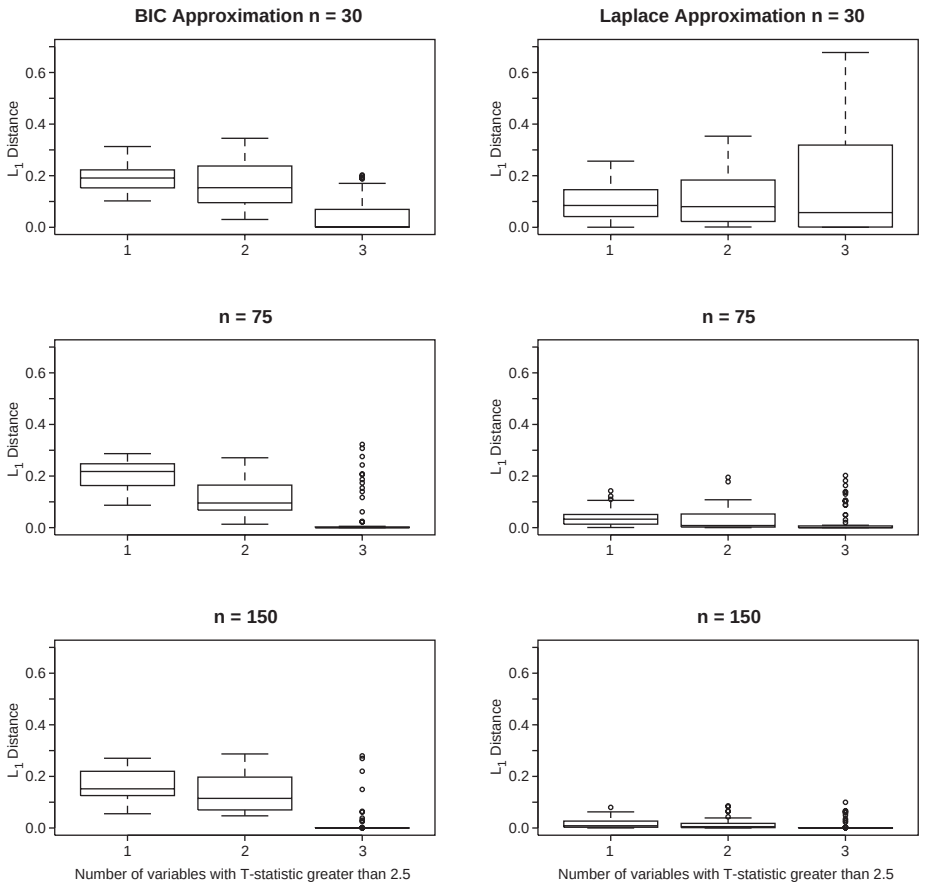

Fig. 2. Box plots of L_1 distances for sample sizes of $n = 30, 75, 150$ for three candidate variables. Distances generated by comparing the BIC approximation to the exact posterior distribution are in the left column and the distances generated by Laplace approximation are in the right column.

significant in the model. Using the samples we obtained the L_1 distance mean of 0.0792 with a standard error of 0.0038. The L_2 and Hellinger distances produced similar results. Thus the same pattern is exhibited with three candidate variables as with five.

The evidence suggests that the BIC approximation's accuracy depends on the number of "significant" variables in the model. This is a very poor property. When some of the variables are not significant the BIC approximation has a serious accuracy problem that persists even in large sample sizes. Since all inferences from BMA are dependent on the posterior model probabilities it is important that these probabilities be accurate to avoid inaccurate inferences. Given this fact, we do not recommend that the BIC approximation be used for BMA.

Table 6
Mean distance between exact posterior distribution and the BIC, Laplace with proper prior and Laplace with Uniform prior approximations to the posterior distribution using random underlying models

n	L_2 Mean			L_1 Mean		
	BIC	Laplace	Uniform	BIC	Laplace	Uniform
50	0.2138	0.1265	0.2898	0.2052	0.1219	0.3025
	(0.0125)	(0.0089)	(0.0153)	(0.0121)	(0.0086)	(0.0158)
100	0.1564	0.0443	0.1871	0.1522	0.0432	0.1917
	(0.0119)	(0.0043)	(0.0137)	(0.0117)	(0.0042)	(0.0139)
150	0.1395	0.0267	0.1610	0.1376	0.0263	0.1635
	(0.0114)	(0.0029)	(0.0130)	(0.0112)	(0.0028)	(0.0132)

Standard errors are shown in parenthesis below the mean.

5. Discussion

While the BIC approximation of Bayes factors may be asymptotically accurate, using this approximation in the context of BMA produces unsatisfactory results. We have illustrated by example how the differences between the methods manifest themselves in practice. Inferences drawn from applications of BMA using the BIC approximation may lead to erroneous conclusions. This paper shows that researchers should be careful when using approximations for BMA. The goal of BMA is to account for the model uncertainty in the analysis. However, when approximations are used to determine the posterior model probabilities, uncertainty associated with approximation accuracy is not accounted for in the inferences.

One of the appealing aspects of using the BIC approximation is that there is no need to specify a prior distribution for the parameters in the model. As an alternative we briefly explored using a uniform prior distribution on the parameters. The uniform prior distribution is an improper prior as is the "unit information" prior. Using a similar approach as above we computed the BIC approximation, Laplace approximation with normal prior and Laplace approximation with uniform prior. Table 6 shows the results of this simulation study. We see that the Laplace approximation with the uniform prior distribution performs worse, on average, than both the BIC approximation and the Laplace approximation with normal prior distribution. Hence, using the "unit information" prior distribution is better than using no, i.e uniform, distribution.

In this paper we have illustrated, via simulation examples and a real example, how exact, Laplace and BIC methods for computing posterior model probabilities compare. The evidence provided in this study suggests that the Laplace approximation seems to perform well. We noticed the accuracy of BIC approximation depends on the number of significant variables in the model. This pattern of inaccuracy holds for large sample sizes as well. In light of this evidence the BIC approximation may not be a good choice in BMA. The Laplace approximation should be the preferred approximation method in the multiple linear regression BMA setting.

Future work includes determining whether the BIC-based approximations for generalized linear models exhibit the same properties. Understanding whether this pattern

persists would guide researchers on the choice of approximation in this setting. Our study is limited by the fact that we only considered the case where the number of candidate regressors is low. Another line of research could include the performance of these methods in the Markov chain Monte Carlo model composition (MC^3) methods. These methods are popular when large number of candidate regressors are present.

References

Chickering, D., Heckerman, D., 1997. Efficient approximations for the marginal likelihood of Bayesian networks with hidden variables. Mach. Learning 29, 181–212.

Clyde, M., 2000. Model uncertainty and health effect studies for particulate matter. Environmetrics 6, 745–763.

Hoeting, J.A., Madigan, D., Raftery, A.E., Volinsky, C.T., 1999. Bayesian model averaging: a tutorial. Statist. Sci. 14, 382–417.

Kass, R.E., Raftery, A.E., 1995. Bayes factors. J. Amer. Statist. Assoc. 90, 773–795.

Kass, R., Wasserman, L., 1995. A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. J. Amer. Statist. Assoc. 90, 928–934.

Lamon, E.C., Clyde, M.A., 2000. Accounting for model uncertainty in prediction of chlorophyll A in lake Okeechobee. J. Agri. Biol. Environ. Statist. 5, 297–322.

Lipkovich, L., 2002. Bayesian model averaging and variable selection in multivariate ecological models. Unpublished Ph.D. Dissertation, Virginia Polytechnic Institute and State University, Blacksburg, VA.

Madigan, D., York, J., 1995. Bayesian graphical models for discrete data. Internat. Statist. Rev. 63, 215–232.

Murphy, M., Wang, D., 2001. Do previous birth interval and maternal education influence infant survival? A Bayesian model averaging analysis of Chinese data. Popul. Stud. 55, 37–48.

Noble, R., 2000. Multivariate applications of Bayesian model averaging. Unpublished Ph.D. Dissertation, Virginia Polytechnic Institute and State University, Blacksburg, VA.

Raftery, A.E., 1996. Approximate Bayes factors and accounting for model uncertainty in generalised linear models. Biometrika 83, 251–266.

Raftery, A.E., 1999. Bayes factors and BIC. Sociological Methods Res. 27, 411–421.

Raftery, A.E., Madigan, D., Hoeting, J.A., 1997. Bayesian model averaging for linear regression models. J. Amer. Statist. Assoc. 92, 179–191.

Raiffa, H., Schlaifer, R., 1961. Applied Statistical Decision Theory. MIT Press, Cambridge, MA.

Schwartz, G., 1978. Estimating the dimension of a model. Ann. Statist. 6, 461–464.

Tierney, L., Kadane, J.B., 1986. Accurate approximations for posterior moments and marginal densities. J. Amer. Statist. Assoc. 81, 82–86.

Viallefont, V., Raftery, A.E., Richardson, S., 2001. Variable selection and Bayesian model averaging in case control studies. Statist. Med. 20, 3215–3230.

Weakliem, D.L., 1999. A critique of the Bayesian information criterion for model selection. Sociological Methods Res. 27, 359–397.