On the Calibration of Bayesian Model Choice Criteria

Pantelis K. Vlachos and Alan E. Gelfand*

Abstract

Model choice is a fundamental problem in data analysis. With interest in hierarchical models which typically arise as Bayesian specifications, we confine ourselves to Bayesian model choice criteria. If $\boldsymbol{Y}$ denotes the observed data and $T(\boldsymbol{Y})$ is the criterion, our goal is to calibrate $T(\boldsymbol{Y})$ in order to assess how large or small it is under a given model. In particular, if we have the distribution of $T(\boldsymbol{Y})$, then we can compute any probabilities or determine any quantities of interest.

Apart from very special cases, analytic development of such distributions is intractable. Standard analytic approximations may be inapplicable if usual random effects are introduced at the various modeling levels. Indeed, calculation of $T(\boldsymbol{Y})$ itself is often difficult enough.

We suggest a generic simulation-intensive approach for obtaining the distribution of $T(\boldsymbol{Y})$ to arbitrary accuracy. We focus on various Bayes factors, e.g., the usual Bayes factor, the posterior Bayes factor and the pseudo-Bayes factor. We illustrate with a binomial regression example.

*Pantelis K. Vlachos is a Research Scientist in the Department of Statistics at Carnegie Mellon University and Alan E. Gelfand is a Professor in the Department of Statistics at the University of Connecticut. The work of the second author was supported in part by NSF grant DMS 99-71206

Key Words: Bayes factor, Monte Carlo, pairwise model comparison, posterior Bayes factor, pseudo-Bayes factor.

1. Introduction

Model choice is a fundamental problem in data analysis and has received considerable attention in the statistical literature. In particular, with the wide availability of inexpensive computing, increasingly complex, hierarchical models are being investigated, making model comparison crucial. In practice, model selection is a utility issue with a suitable criterion adopted for a given setting reflecting the intended use for the model in that setting. For nested models, model choice might be viewed as hypothesis testing but, in general, it must reside in predictive performance since this is the only space in which all models can be compared.

In this regard, a considerable number of criteria have been proposed. It is not our objective here to add to this collection nor to enter into the philosophical debate regarding which, among existing ones, is preferred. Rather, having identified an appropriate criterion, our goal is to attempt to calibrate this criterion in order to assess how large or small it is under a given model.

Since the previously mentioned models are typically Bayesian specifications, we confine ourselves to Bayesian model choice criteria. More precisely, if $\boldsymbol{Y}$ denotes the observed data and $T(\boldsymbol{Y})$ is the criterion, we seek the distribution of $T(\boldsymbol{Y})$ under a particular model. In application, obtaining the distribution of $T(\boldsymbol{Y})$ is hopeless except for very special cases. Indeed, calculation of $T(\boldsymbol{Y})$ itself is often difficult enough. For certain T, in certain cases, the approximate distribution for $T(\boldsymbol{Y})$ can be obtained. Such an approximation depends upon regularity conditions which do not hold for general multilevel models incorporating random effects.

We suggest a generic simulation-intensive approach for obtaining the distribution of $T(\boldsymbol{Y})$ to arbitrary accuracy. Since T is one-dimensional, we can implement a Monte Carlo testing approach, generating a sample of univariate T^* under a given model to then see where our observed $T(\boldsymbol{Y})$ falls, to enable calibration of the observed $T(\boldsymbol{Y})$ under the model. The strategy becomes simulation-intensive due to the repeated computation of the T^*'s. Such computation requires repeated generation of data under the given model. In this sense the Bayesian model must be proper, i.e., only proper priors can be utilized.

The most widely used Bayesian model choice criterion is the Bayes factor. See Kass and Raftery (1995) for an extensive discussion and references. The Bayes factor compares models in pairs and rough calibration has been provided, dating to Jeffreys (1961, Appendix B) who suggests interpretation in half-units on the $\log_{10}$ scale. In comparing model M_1 against model M_2, denoting the Bayes factor by B_{12}, we have

$\log_{10}(B_{12})$	Evidence against model M_2
0 to 0.5	not worth more than a bare mention
0.5 to 1	substantial
1 to 2	strong
> 2	decisive

Table 1: Jeffreys' scale of evidence

Since $\sqrt{10} = 3.17$, a Bayes factor between 1/3 and 3 should be considered to provide no real evidence to distinguish the models. It would be useful, in practice to be able to compute

$\Pr(\log_{10} B_{12} > c|M_i)$, $i = 1, 2$, in order to quantify Jeffreys' interpretation. Recent attempts to quantify "surprise" appear in Bayarri and Berger (1999) and in Sellke, Bayarri and Berger (1999) though they do not address the distribution of B_{12}.

Alternatives to the Bayes factor have been proposed; using cross-validation ideas the pseudo-Bayes factor arises (Geisser and Eddy, 1979). Aitkin (1991) proposed the posterior Bayes factor, Berger and Pericchi (1992) introduced the intrinsic Bayes factor while O'Hagan (1995) suggested the use of fractional Bayes factors. Recently, Gelfand and Ghosh (1998) proposed a class of posterior predictive criteria. Again, while these criteria decide between two models, one is frequently left wondering how different the models are. Attaching probabilities to these differences under the respective models would provide valuable clarification.

The organization of the paper is as follows. In section 2 we formalize notation and our objective. In section 3 we discuss briefly cases when exact and asymptotic distribution theory is available. In section 4 we detail our approach. Finally in section 5 we present an illustrative example.

2. The objective

We focus on pairwise comparison of models offering a remark on simultaneous comparison of k models at the end of the section. Assuming an $n \times 1$ data vector $\boldsymbol{Y}$, the ingredients include the Bayesian models with respective first stage distributions $f(\boldsymbol{Y}|\boldsymbol{\theta}_1^{(i)}, M_i)$ and priors $\pi(\boldsymbol{\theta}_1^{(i)}|M_i)$, $i = 1, 2$. In fact, the priors are customarily specified hierarchically so that $\pi(\boldsymbol{\theta}_1^{(i)}|M_i)$ is often not available explicitly. We do not attempt a more general representation for $\pi(\boldsymbol{\theta}_1^{(i)}|M_i)$; we only require that it is proper. We also recognize that other marginalizations are possible, yielding alternative "first stage likelihoods" but since we work in predictive space, i.e., the space of the $\boldsymbol{Y}$'s, marginalizing over all parameters, this is not of consequence. Typically, the first stage (and sometimes the prior) involves covariates. These are taken as fixed and are suppressed in our notation.

The remaining ingredient is a criterion $T(\boldsymbol{Y})$ which, without loss of generality, can always be designed (centered) such that if $T(\boldsymbol{Y}) > 0$ we choose model M_1 while if $T(\boldsymbol{Y}) < 0$ we choose model M_2. We seek the distribution of $T(\boldsymbol{Y})$ under M_i, $i = 1, 2$. This will not only enable computation of e.g., $\Pr(T(\boldsymbol{Y}) > 0|M_1)$ and $\Pr(T(\boldsymbol{Y}) < 0|M_2)$ but generally $\Pr(T(\boldsymbol{Y}) > c|M_i)$, quantiles of $T(\boldsymbol{Y})$ under M_i and $\Pr(T(\boldsymbol{Y}) > T(\boldsymbol{Y}_{obs})|M_i)$ where $\boldsymbol{Y}_{obs}$ denotes the observed data vector.

We focus on the Bayes factor (BF), the posterior Bayes factor ($PoBF$) due to Aitkin (1991), and the pseudo-Bayes factor ($PsBF$) dating to Geisser and Eddy (1979), in our illustrations. In particular,

$$BF_{12} = \frac{f(\boldsymbol{Y}_{obs}|M_1)}{f(\boldsymbol{Y}_{obs}|M_2)}, \tag{2·1}$$

where $f(\boldsymbol{Y}|M_i) = \int f(\boldsymbol{y}|\boldsymbol{\theta}_1^{(i)}, M_i)\pi(\boldsymbol{\theta}_1^{(i)}|M_i)d\boldsymbol{\theta}_1^{(i)}$. To align with Jeffreys' scale we take $T(\boldsymbol{Y}) = \log_{10} BF_{12} (= 0.4343 \ln BF_{12})$. Next

$$PoBF_{12} = \frac{f(\boldsymbol{Y}_{obs}|\boldsymbol{Y}_{obs}, M_1)}{f(\boldsymbol{Y}_{obs}|\boldsymbol{Y}_{obs}, M_2)}, \tag{2·2}$$

where $f(\boldsymbol{Y}|\boldsymbol{Y}_{obs}, M_i) = \int f(\boldsymbol{Y}|\boldsymbol{\theta}_1^{(i)}, \boldsymbol{Y}_{obs}, M_i)\pi(\boldsymbol{\theta}_1^{(i)}|\boldsymbol{Y}_{obs}, M_i)d\boldsymbol{\theta}_1^{(i)}$. With conditionally independent first stage observations, $\boldsymbol{Y}_{obs}$ does not appear in the first density under the integral.

Here, we take $T(\boldsymbol{Y}) = \ln(PoBF_{12})$. Finally,

$$PsBF_{12} = \frac{\prod_{j=1}^{n} f(Y_{j,obs}|\boldsymbol{Y}_{(j),obs}, M_1)}{\prod_{j=1}^{n} f(Y_{j,obs}|\boldsymbol{Y}_{(j),obs}, M_2)}, \tag{2·3}$$

where $\boldsymbol{Y}_{(j)}$ denotes $\boldsymbol{Y}$ with Y_j deleted and $f(Y_j|\boldsymbol{Y}_{(j),obs}, M_i) =$ $\int f(Y_j|\boldsymbol{Y}_{(j),obs}, \boldsymbol{\theta}_1^{(i)}, M_i)\pi(\boldsymbol{\theta}_1^{(i)}|\boldsymbol{Y}_{(j),obs}, M_i)d\boldsymbol{\theta}_1^{(i)}$. With conditionally independent first stage observations, $\boldsymbol{Y}_{(j),obs}$ does not appear in the first density under the integral. Here, we take $T(\boldsymbol{Y}) = \ln(PsBF_{12})$. No qualitative calibration, analogous to Jeffreys', is available for $PoBF$ or $PsBF$.

Computation of BF has been a topic of considerable recent discussion in the literature. It may be calculated through the computation of the $f(\boldsymbol{Y}_{obs}|M_i)$ individually (see, e.g., Newton and Raftery, 1994; Gelfand and Dey (1994); Chib (1995); Chib and Jeliaskov, 1999) or the ratio directly (see e.g., Spiegelhalter and Smith, 1982; Verdinelli and Wasserman, 1995; Meng and Wong, 1996; DiCiccio, Kass, Raftery and Wasserman, 1997). Computation of $PoBF$ is routine nowadays. Customary Markov chain Monte Carlo model fitting provides posterior samples under each model, hence straightforward Monte Carlo integration for the numerator and denominator in (2·2). Computation of $PsBF$ requires the cross-validation predictive densities $f(Y_j|\boldsymbol{Y}_{(j)}, M_i)$. A weighted bootstrap can be used here as illustrated in Gelfand, Dey and Chang (1992). Gelfand and Dey (1994) provide general simulation strategies for calculation of all Bayes factors mentioned in the introduction. Again, computation of any of these Bayes factors is apart from developing their distribution under a given model.

Finally, if we seek to compare k models simultaneously, we would introduce a so-called "screening" criterion (Kass and Raftery, 1995) which is computed for each model. Without loss of generality, the criterion can be designed so that the model achieving the largest value is selected. For instance, the criterion could be $f(\boldsymbol{Y}_{obs}|M_i)$, $f(\boldsymbol{Y}_{obs}|\boldsymbol{Y}_{obs}, M_i)$ or $\prod_{j=1}^{n} f(Y_{j,obs}|\boldsymbol{Y}_{(j),obs}, M_i)$. Denoting the criterion computed under M_i by $S(\boldsymbol{Y}; M_i)$ we would require the k^2 distributions $S(\boldsymbol{Y}; M_i)|M_j$ for $i = 1, \ldots, k$, $j = 1, \ldots, k$. Obtaining all of these distributions will become computationally prohibitive and, even if we did, interpretation will become very difficult. The advantage of simplification to one criterion studied under just two models is evident.

3. Analytic Results

We begin with an elementary example where exact calculations are straightforward. Suppose $Y_1, \ldots, Y_n$ are a sample which, under M_1 is $N(\theta_o, \sigma^2)$, σ^2 known. Under M_2, $Y_1, \ldots, Y_n$ are a sample from $N(\theta, \sigma^2)$ with $\theta \sim N(0, \tau^2)$, again σ^2 known as well as τ^2. We can reduce to $\bar{Y}$ by sufficiency and note that $f(\bar{Y}|M_1) = N(\theta_o, \sigma^2/n)$ while $f(\bar{Y}|M_2) = N(0, \frac{\sigma^2}{n} + \tau^2)$. Hence,

$$\ln BF_{12} = \frac{1}{2}\ln(1 + a_n) - \left[\frac{(\bar{Y} - \theta_o)^2}{2\sigma^2/n} - \frac{\bar{Y}^2}{2\sigma^2(1 + a_n)/n}\right]$$

where $a_n = n\tau^2/\sigma^2$. After a little algebra, we obtain

$$-2\ln BF_{12} + \ln(1 + a_n) + \frac{\theta_o^2}{\tau^2} = \frac{n}{\sigma^2}\frac{a_n}{1 + a_n}\left(\bar{Y} - \frac{(1 + a_n)\theta_o}{a_n}\right)^2.$$

Now under M_1, $\left(\bar{Y} - \frac{(1+a_n)\theta_o}{a_n}\right)^2 \sim \frac{\sigma^2}{n}\chi^2_{1,\theta_o^2/2\tau^2 a_n}$ while under M_2, $\left(\bar{Y} - \frac{(1+a_n)\theta_o}{a_n}\right)^2 \sim \frac{\sigma^2}{n}(1 + a_n)\chi^2_{1,\theta_o^2 n(1+a_n)/2\sigma^2 a_n^2}$. Thus, with $c_n = \ln(1 + a_n) + \frac{\theta_o^2}{\tau^2}$,

$$-2\ln BF_{12}+c_n \sim \frac{a_n}{1+a_n}\chi^2_{1,\theta_o^2/2\tau^2 a_n} \text{ under } M_1$$

$$-2\ln BF_{12}+c_n \sim a_n\chi^2_{1,\theta_o^2 n(1+a_n)/2\sigma^2 a_n^2} \text{ under } M_2. \tag{3·1}$$

From (3·1), we have the exact distribution of BF_{12} under each model. Densities can be formally plotted to make comparison. $\Pr(BF_{12} > b|M_1) = \Pr\left(Z < \frac{1+a_n}{a_n}(c_n - 2\ln b)|Z \sim \chi^2_{1,\theta_o^2/2\tau^2 a_n}\right)$. Similarly, $\Pr(BF_{12} < b|M_2) = \Pr\left(Z > (c_n - 2\ln b)/a_n|Z \sim \chi^2_{1,\theta_o^2 n(1+a_n)/2\sigma^2 a_n^2}\right)$.

Next, $f(\bar{Y}_{new}|\bar{Y}_{obs}, M_1) = N(\theta_o, \sigma^2/n)$ but $f(\bar{Y}_{new}|\bar{Y}_{obs}, M_2) = N(\tilde{a}_n\bar{Y}_{obs}, (1+\tilde{a}_n)\sigma^2/n)$ where $\tilde{a}_n = n\tau^2/(\sigma^2+n\tau^2)$. Hence, mimicking calculations which led to (3·1), we can connect $PoBF_{12}$ to a non-central chi-square random variable under M_1, to a different one under M_2.

Also, $f(Y_j|\boldsymbol{Y}_{(j)}, M_1) = N(\theta_o, \sigma^2)$ while $f(Y_j|\boldsymbol{Y}_{(j)}, M_2) = N(\tilde{a}_{n-1}\bar{Y}_{(j)}, (1+\tilde{a}_{n-1}/(n-1))\sigma^2)$. After insertion into (2·3) and, again, mimicking calculations which led to (3·1), we can also connect $PsBF_{12}$ to non-central chi-square random variables under M_1 and M_2.

As an illustration, let $n = 20$, $\theta_o = 0$, $\sigma^2 = 1$ and $\tau^2 = 5$. The distributions of the $\log_{10} BF_{12}$, $\ln PoBF_{12}$ and $\ln PsBF_{12}$ are shown in Figure 1, while in Table 2 we present probabilities of interest for the half-units in Jeffreys' scale of evidence. The figure shows that the scale of the BF is roughly half of that for $PoBF$ and $PsBF$. Qualitative calibration using Table 1 is inappropriate for the latter pair. Perhaps more important is the severe asymmetry in the strength of evidence between corresponding pairs of distributions. For

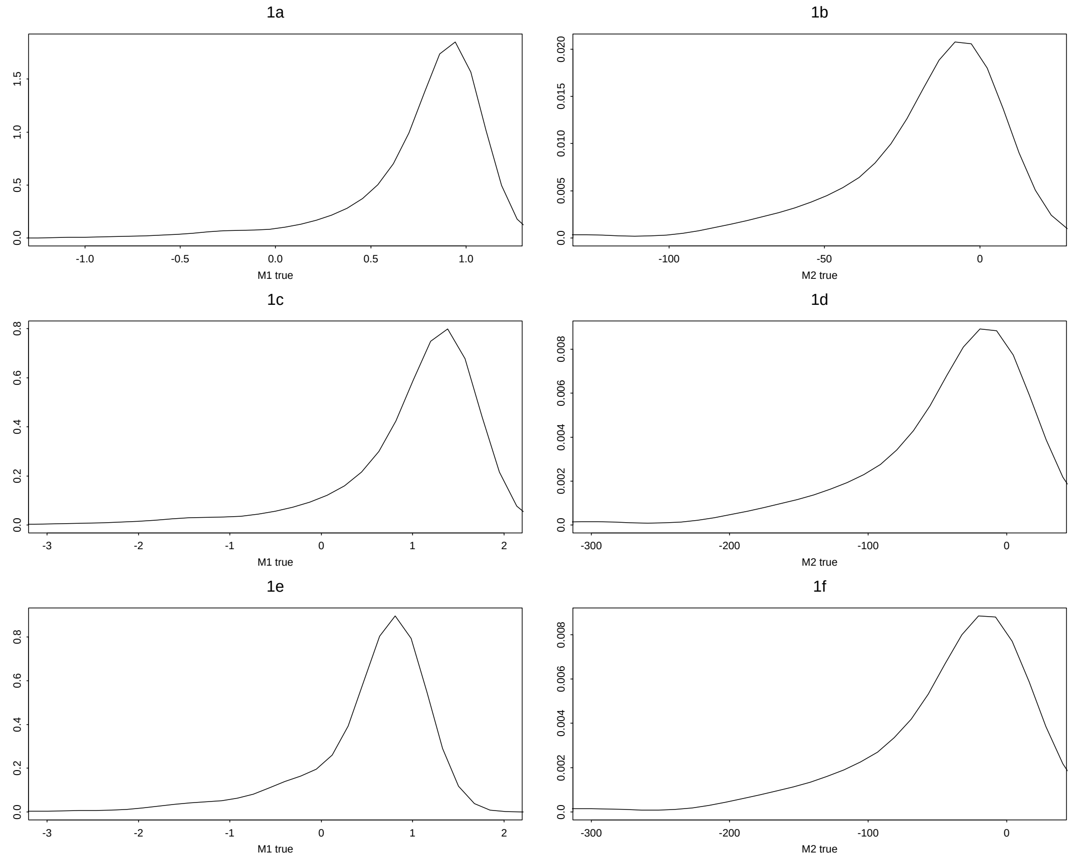

Figure 1: The densities of various Bayes factors for testing a precise null hypothesis. See text for details. 1a) is $\log_{10} BF_{12}$ under M_1, 1b) is $\log_{10} BF_{12}$ under M_2, 1d) is $\ln PoBF_{12}$ under M_1, 1d) is $\ln PoBF_{12}$ under M_2, 1e) is $\ln PsBF_{12}$ under M_1, 1f) is $\ln PsBF_{12}$ under M_2

instance, using BF, a finding of strong or decisive evidence for model 1 under M_1 is rather unlikely (probability $= 0.07$) while a finding of strong or decisive evidence for model 2 under M_2 is very likely (probability $= 0.75$).

More generally, consider the nested normal linear model situation with $\boldsymbol{Y} = X_1\boldsymbol{\beta}_1 + \boldsymbol{\epsilon}$ under M_1, $\boldsymbol{Y} = X_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}$ under M_2. Here $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\boldsymbol{I})$ with again σ^2 known, $\boldsymbol{X}_1^T\boldsymbol{X}_2 = \mathbf{0}$ (for computational convenience) and $\pi(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) = N(\boldsymbol{\beta}_1|\mathbf{0}, \sigma^2\boldsymbol{V}_1)\cdot N(\boldsymbol{\beta}_2|\mathbf{0}, \sigma^2\boldsymbol{V}_2)$ (again for computational simplicity). Then $-2\ln BF_{12} + c_n = \sigma^2 Z$ where $Z = \hat{\boldsymbol{\beta}}_2^T\boldsymbol{X}_2^T\boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2$ with $\hat{\boldsymbol{\beta}}_2$ the usual least squares estimator and c_n depends upon $|(\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1} + \boldsymbol{V}_1^{-1}|$, $|(\boldsymbol{X}_2^T\boldsymbol{X}_2)^{-1} + \boldsymbol{V}_2^{-1}|$,

	c			
	0	0.5	1	2
$\Pr(\log_{10} BF_{12} > c\|M_1)$	0.958	0.862	0.074	0.000
$\Pr(\log_{10} BF_{21} > c\|M_2)$	0.828	0.782	0.750	0.696
$\Pr(\ln PoBF_{12} > c\|M_1)$	0.908	0.830	0.692	0.000
$\Pr(\ln PoBF_{21} > c\|M_2)$	0.860	0.838	0.816	0.782
$\Pr(\ln PsBF_{12} > c\|M_1)$	0.828	0.704	0.138	0.000
$\Pr(\ln PsBF_{21} > c\|M_2)$	0.888	0.860	0.838	0.802

Table 2: Probabilities of interest for various Bayes factors

$|\boldsymbol{V}_1|$ and $|\boldsymbol{V}_2|$. See Smith and Spiegelhalter (1980, p214) for details. Again BF_{12} can be connected to chi-square random variables under M_1 and M_2. The form for $-2\ln BF_{12}$ when $\boldsymbol{X}_1^T\boldsymbol{X}_2 \neq \boldsymbol{0}$ is given in Lempers (1971, p35). The form for non-nested M_1 and M_2 is given in Atkinson (1978). In this latter case, BF_{12} can only be connected to a linear combination on non-central chi-squared random variables, a much less attractive distributional situation.

Similar results apply for $PoBF_{12}$ since the posteriors for $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)^T$ and for $\boldsymbol{\beta}_1$, which will be inserted into (2·2), will be normal as in the BF_{12} case. We omit details. However, the cross-validatory form for $PsBF_{12}$ does not allow such simplification.

In the case when σ^2 is unknown but inverse Gamma priors are used, again exact calculations for BF_{12} can be carried out, obtaining functions of the model and residual sum of squares for each model. In the nested case, under certain limits, BF_{12} can be connected to the usual F statistic (Lempers, 1971). Similarly, Aitkin (1991) shows that, under certain limits, $PoBF_{12}$ can be connected to the ratio of residual sums of squares. Again, in the nested case, we connect with the usual F statistic. Lastly, Geisser(1993) provides analogous approximate calculations for $\prod_{j=1}^{n} f(y_j|\boldsymbol{y}_{(j)})$ under normal linear models. Again, the cross-validation form precludes connection to the ratio of the residual sums of squares.

In the non-Gaussian case, apart from specifications exhibiting special conjugacies, exact calculation of the various Bayes factors is intractable. Under appropriate conditions, the Laplace approximation (Tierney and Kadane, 1986) can be employed, as illustrated for the Bayes factor in Kass and Vaidyanathan (1992). Asymptotics for the posterior Bayes factor are given in Aitkin (1991) and for the pseudo-Bayes factor in Stone (1977). A summary of all of these asymptotics is presented in Gelfand and Dey (1994).

In the nested case, these approximations connect the log of the various Bayes factors to $-2\log\lambda$ where λ denotes the usual likelihood ratio statistic. More precisely $T(\boldsymbol{Y}) = -2\log\lambda + h_n(\boldsymbol{Y})$ where, for large n, $h_n(\boldsymbol{Y})$ is roughly a computable constant under M_1 and M_2. Under familiar regularity conditions, $-2\log\lambda$ tends to a chi-square distribution implying a simple approximation to the distribution of $T(\boldsymbol{Y})$. However the accuracy of this approximation is unknown in practice. Rather, it is used typically, to clarify asymptotic behavior of $T(\boldsymbol{Y})$ with regard to, e.g., consistency and Lindley's paradox (Lindley, 1957). It also provides a penalty to the full model relative to the reduced with regard to dimension. Unfortunately, such approximation is inapplicable to models where dimension increases with sample size, as in most random effects models. We seek the exact distribution of $T(\boldsymbol{Y})$.

4. Exact distribution theory using simulation

If we can simulate random realizations of $T(\boldsymbol{Y})$ either under M_1 or under M_2, then we can learn, with arbitrary accuracy, about the distribution of $T(\boldsymbol{Y})$ in each case. However, given say under model M_1, suppose we generate $\boldsymbol{Y}^*_{rep,b}$, $b = 1, \ldots, B$, each viewed as a replicate of $\boldsymbol{Y}_{obs}$. That is we generate $\boldsymbol{\theta}^{(1)*}_{1,b} \sim \pi(\boldsymbol{\theta}^{(1)}_1 | M_1)$ and then $\boldsymbol{Y}^*_{rep,b}$ from $f(\boldsymbol{Y}|\boldsymbol{\theta}^{(1)*}_{1,b}, M_1)$, incorporating whatever covariates were associated with the original $\boldsymbol{\theta}_1^{(1)}$ and the observed $\boldsymbol{Y}$. Therefore $t_b^* = T(\boldsymbol{Y}_{rep,b}^*) \sim T(\boldsymbol{Y})|M_1$. Similarly, we can obtain samples from $T(\boldsymbol{Y})|M_2$.

In fact, if we can calculate $T(\boldsymbol{Y}_{obs})$ then presumably we can calculate $T(\boldsymbol{Y}_{rep,b}^*)$ for a given $\boldsymbol{Y}_{rep,b}^*$. Such calculation can be time-consuming. For instance, for the $PoBF$ or the $PsBF$ we require posterior samples to obtain the proposed Monte Carlo integrations. Such sampling will in turn require a Markov chain Monte Carlo model fitting. Hence, we will need $B-$many MCMC model fittings to obtain t_b^*, $b = 1, \ldots, B$ under model M_i. Once obtained, these samples can be used to create density estimates, to calculate desired probabilities and to obtain quantiles of interest. We illustrate in the next section.

5. An illustrative example

We illustrate our method with the use of a random effects binomial regression model. We use a dataset abstracted from the United Network for Organ Sharing (UNOS) Public-Use database. The overall database consists of 3,688 transplants at 131 transplant centers during the period October 1, 1987 to December 31, 1989. We confine ourselves to the 10 centers with the largest number of heart transplants. We classify subjects into age groups centered at 25, 35, 45, 55 and 65 years of age and record a 1 or 0 according to whether or not the patient developed short term problems leading to rejection of the organ or death. We obtained the data set exhibited in Table 3.

To model the data, we begin with M_1, a binomial model for the n_{ij} patients in the j^{th} age group at the i^{th} center, i.e., y_{ij}, the number of patients developing problems, is distributed as $Bin(n_{ij}, p_{ij})$, with $i = 1, \ldots, I = 10$ and $j = 1, \ldots, J = 5$.

Age Group						
Center	25	35	45	55	65	Total
1	0/3	0/11	9/27	16/36	0/2	25/79
2	3/7	1/8	4/15	10/31	7/15	25/76
3	6/15	3/11	2/14	7/30	0/5	18/75
4	1/14	1/6	2/23	10/32	5/16	19/91
5	3/14	1/6	4/19	5/28	0/7	13/74
6	2/11	1/12	3/23	12/51	3/9	21/106
7	5/16	6/19	5/34	9/39	11/26	36/134
8	1/9	2/7	3/28	4/25	0/4	10/73
9	6/33	7/17	12/42	11/42	4/18	40/152
10	13/31	2/7	5/28	8/34	0/1	28/101

Table 3: For UNOS Heart transplant data, for 10 centers and 5 age groups, number of patients developing problems leading to short term rejection of organ or death/number of transplant patients.

To explain the p_{ij} using age, we propose a logistic regression model at the hospital level, $\log\left(p_{ij}/(1-p_{ij})\right) = \alpha_i + \beta_i x_j$, where the ages, x_j have been centered and scaled. To complete the Bayesian specification, we assume at the second stage exchangeable (α_i, β_i) such that

$$\begin{pmatrix} \alpha_i \\ \beta_i \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_\alpha \\ \mu_\beta \end{pmatrix}, \begin{pmatrix} \tau_\alpha^2 & 0 \\ 0 & \tau_\beta^2 \end{pmatrix}\right)$$

with μ_α and μ_β being independent normal a priori and τ_α^2, τ_β^2 being independent inverse Gamma a priori. Prior information suggests that roughly 30% of transplant patients develop problems and that oldest patients have roughly twice the chance of youngest ones. After a bit of algebra this suggests μ_α centered about $m_\alpha = -.9$, μ_β centered about $m_\beta = .17$. As for the prior variance of μ_α we consider two informative choices: $v_\alpha^2 = (.2)^2$ which keeps the p_{ij} between .18 and .42 at $\beta = 0$ and $v_\alpha^2 = (.13)^2$ which keeps the p_{ij} between .21 and .38

at $\beta = 0$. For the prior variance of μ_β we consider $v_\beta^2 = (.05)^2$ which is informative to the extent that it essentially insures $\mu_\beta > 0$. For τ_α^2 we assume an inverse Gamma distribution, $IGam(c, d)$ with mean $(.2)^2$ and standard deviation .1; for τ_β^2 an inverse Gamma, $IGam(e, f)$ with mean $(.05)^2$ and standard deviation .03. These priors are arbitrary but attempt to suggest a realistic amount of variability for the α_i's and β_i's.

As an alternative model (M_2), we will consider one that does not include a covariate for the patient ages. We let $p_{ij} = \exp\{\alpha_i\}/(1 + \exp\{\alpha_i\})$, where $\alpha_i \sim N(\mu_\alpha, \tau_\alpha^2)$ and as before $\mu_\alpha \sim N(m_\alpha, v_\alpha^2)$ (m_α and v_α^2 set as in the full model), and for τ_α^2 we assume an inverse Gamma distribution with mean $(.2)^2$ and standard deviation .1. Thus, M_2 is nested within M_1.

We calculate all of the Bayes factors using draws of variates from the posterior distributions $\pi(\alpha_i, \beta_i, \mu_\alpha, \mu_\beta, \tau_\alpha^2, \tau_\beta^2 | \boldsymbol{y}, M_1)$ and $\pi(\alpha_i, \mu_\alpha, \tau_\alpha^2 | \boldsymbol{y}, M_2)$, following Gelfand and Dey (1994). For both models, the Metropolis-Hastings algorithm is used to obtain the posterior samples. The complete conditional distributions, using the notation of Gelfand and Smith (1990), are

$$
\begin{aligned}
[\alpha_i | \beta_i, \mu_\alpha, \mu_\beta, \tau_\alpha^2, \tau_\beta^2, \boldsymbol{y}] &\propto \frac{N\left(\alpha_i | \mu_\alpha + \tau_\alpha^2 \sum_{j=1}^J y_{ij}, \tau_\alpha^2\right)}{\prod_{j=1}^J (1 + \exp\{\alpha_i + \beta_i x_j\})^{n_{ij}}} \\
[\beta_i | \alpha_i, \mu_\alpha, \mu_\beta, \tau_\alpha^2, \tau_\beta^2, \boldsymbol{y}] &\propto \frac{N\left(\beta_i | \mu_\beta + \tau_\beta^2 \sum_{j=1}^J y_{ij}, \tau_\beta^2\right)}{\prod_{j=1}^J (1 + \exp\{\alpha_i + \beta_i x_j\})^{n_{ij}}} \\
[\mu_\alpha | \alpha_i, \beta_i, \mu_\beta, \tau_\alpha^2, \tau_\beta^2, \boldsymbol{y}] &\sim N\left(\mu_\alpha | \frac{v_\alpha^2 \sum_{i=1}^I \alpha_i + \tau_\alpha^2 m_\alpha}{v_\alpha^2 + \tau_\alpha^2}, \frac{v_\alpha^2 \tau_\alpha^2}{v_\alpha^2 + \tau_\alpha^2}\right) \\
[\mu_\beta | \alpha_i, \beta_i, \mu_\alpha, \tau_\alpha^2, \tau_\beta^2, \boldsymbol{y}] &\sim N\left(\mu_\beta | \frac{v_\beta^2 \sum_{i=1}^I \beta_i + \tau_\beta^2 m_\beta}{v_\beta^2 + \tau_\beta^2}, \frac{v_\beta^2 \tau_\beta^2}{v_\beta^2 + \tau_\beta^2}\right) \\
[\tau_\alpha^2 | \alpha_i, \beta_i, \mu_\alpha, \mu_\beta, \tau_\beta^2, \boldsymbol{y}] &\sim IGam\left(\frac{I}{2} + c, \left[\frac{\sum_{i=1}^I (\alpha_i - \mu_\alpha)^2}{2} + \frac{1}{d}\right]^{-1}\right) \\
[\tau_\beta^2 | \alpha_i, \beta_i, \mu_\alpha, \mu_\beta, \tau_\alpha^2, \boldsymbol{y}] &\sim IGam\left(\frac{I}{2} + e, \left[\frac{\sum_{i=1}^I (\beta_i - \mu_\beta)^2}{2} + \frac{1}{f}\right]^{-1}\right)
\end{aligned}
$$

for M_1, and

$$
\begin{aligned}
[\alpha_i|\mu_\alpha, \tau_\alpha^2, \boldsymbol{y}] &\propto \frac{N\left(\alpha_i|\mu_\alpha + \tau_\alpha^2 \sum_{j=1}^{J} y_{ij}, \tau_\alpha^2\right)}{\prod_{j=1}^{J}\left(1+\exp\{\alpha_i\}\right)^{n_{ij}}} \\
[\mu_\alpha|\alpha_i, \tau_\alpha^2, \boldsymbol{y}] &\sim N\left(\mu_\alpha|\frac{v_\alpha^2 \sum_{i=1}^{I} \alpha_i + \tau_\alpha^2 m_\alpha}{v_\alpha^2 + \tau_\alpha^2}, \frac{v_\alpha^2 \tau_\alpha^2}{v_\alpha^2 + \tau_\alpha^2}\right) \\
[\tau_\alpha^2|\alpha_i, \mu_\alpha, \boldsymbol{y}] &\sim IGam\left(\frac{I}{2} + c, \left[\frac{\sum_{i=1}^{I}(\alpha_i - \mu_\alpha)^2}{2} + \frac{1}{d}\right]^{-1}\right)
\end{aligned}
$$

for M_2.

The logarithm of the Bayes factor of M_1 over M_2 is equal to 0.632 for the data given, which according to Jeffreys scale of evidence (Table 1) reflects substantial evidence against M_2. The distributions of the logarithm of the Bayes factor under M_1 and M_2 are shown in the form of histograms in Figure 2. In particular we can summarize the findings by calculating the estimated probabilities of each one of the categories in Jeffreys scale of evidence. Table 4 shows that even when the full model (M_1) is true, the Bayes factor would select the reduced model (M_2) more than 50% of the time (0.525). On the other hand, when the reduced model is true, the Bayes factor picks the correct model 91% of the time. The observed Bayes factor value of 0.632 is clearly within the range of plausible values when M_1 is true, but it is nearly "extreme" when M_2 is true ($\Pr(\log_{10}(BF_{12}) > 0.632|M_2) \approx 0.051$), i.e., at best, suggestively "surprising" under M_2.

The distributions of the $\ln PoBF_{12}$ and $\ln PsBF_{12}$ under M_i, $i = 1, 2$ are presented in Figures 3 and 4 respectively. Note again, the difference in scale relative to that of BF_{12}.

	Evidence against	
	M_1	M_2
< 0	0.525	0.089
$0 - 0.5$	0.103	0.048
$0.5 - 1$	0.082	0.091
$1 - 2$	0.149	0.229
> 2	0.141	0.543

Table 4: Estimated categorical probabilities using Jeffreys' scale of evidence for the heart transplant data

References

Aitkin, M. (1991). Posterior Bayes factors (with discussion). *Journal of the Royal Statistical Society, Series B* **53**, 111–128.

Atkinson, A. C. (1978). Posterior probabilities for choosing a regression model. *Biometrika* **65**, 39–48.

Bayarri, M. J. & Berger, J. O. (1999). Quantifying surprise in the data and model verification (with discussion). In *Bayesian Statistics 6. Proceedings of the Sixth Valencia International Meeting.* Oxford University Press, Oxford, 53–82.

Berger, J. O. & Pericchi, L. R. (1996). The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association* **91**, 109–122.

Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association* **90**, 1313–1321.

Chib, S. & Jeliaskov, I. (1999). Marginal likelihood from the Metropolis-Hastings output. Technical report, Olin School of Business, Washington University of St. Louis.

DiCiccio, T. J., Kass, R. E., Raftery, A. E., & Wasserman, L. (1997). Computing Bayes factors by combining simulation and asymptotic approximations. *Journal of the*

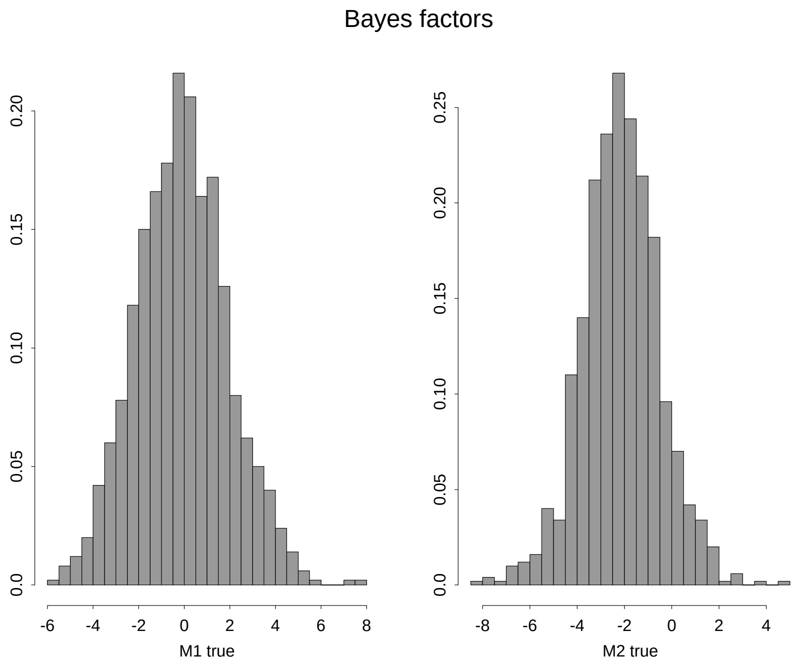

Figure 2: Distributions of $\log_{10}(BF_{12})$ under M_1 and M_2 for UNOS data

American Statistical Association **92**, 903–915.

Geisser, S. (1993). *Predictive Inference. An Introduction.* Chapman & Hall, London.

Geisser, S. & Eddy, W. F. (1979). A predictive approach to model selection (corr: V75 p765). *Journal of the American Statistical Association* **74**, 153–160.

Gelfand, A. E. & Dey, D. K. (1994). Bayesian model choice: Asymptotics and exact calculations. *Journal of the Royal Statistical Society, Series B* **56**, 501–514.

Gelfand, A. E., Dey, D. K., & Chang, H. (1992). Model determination using predictive distributions, with implementation via sampling-based methods (with discussion). In *Bayesian Statistics 4. Proceedings of the Fourth Valencia International Meeting.* Clarendon Press, Oxford, 147–159.

Gelfand, A. E. & Ghosh, S. K. (1998). Model choice: A minimum posterior predictive loss approach. *Biometrika* **85**, 1–11.

Posterior Bayes factors

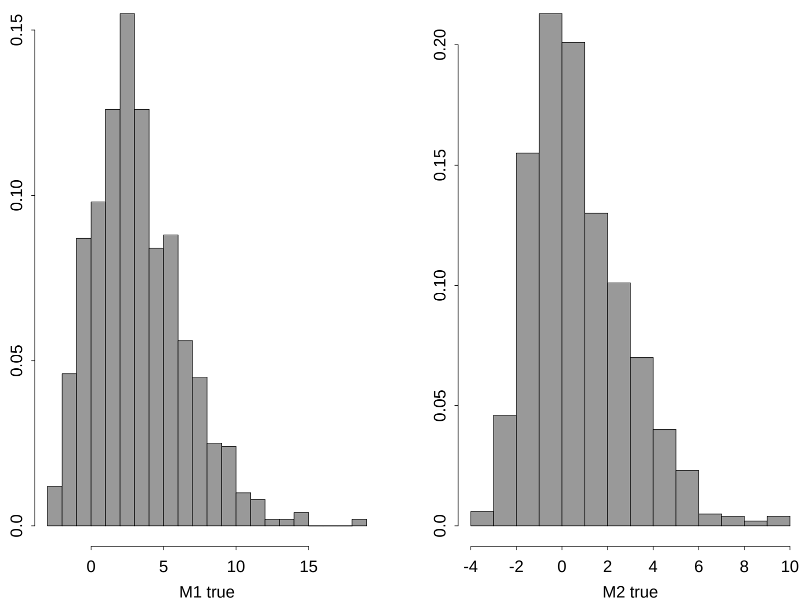

Figure 3: Distributions of $\ln(PoBF_{12})$ under M_1 and M_2 for UNOS data

Gelfand, A. E. & Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association* **85**, 398–409.

Jeffreys, H. (1961). *Theory of Probability.* Oxford University Press, London.

Kass, R. E. & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association* **90**, 773–795.

Kass, R. E. & Vaidyanathan, S. K. (1992). Approximate Bayes factors and orthogonal parameters, with application to testing equality of two binomial proportions. *Journal of the Royal Statistical Society, Series B* **54**, 129–144.

Lempers, F. B. (1971). *Posterior Probabilities of Alternative Linear Models.* University Press, Rotterdam.

Lindley, D. V. (1957). A statistical paradox. *Biometrika* **44**, 187–192.

Meng, X-L & Wong, W. H. (1996). Simulating ratios of normalizing constants via a

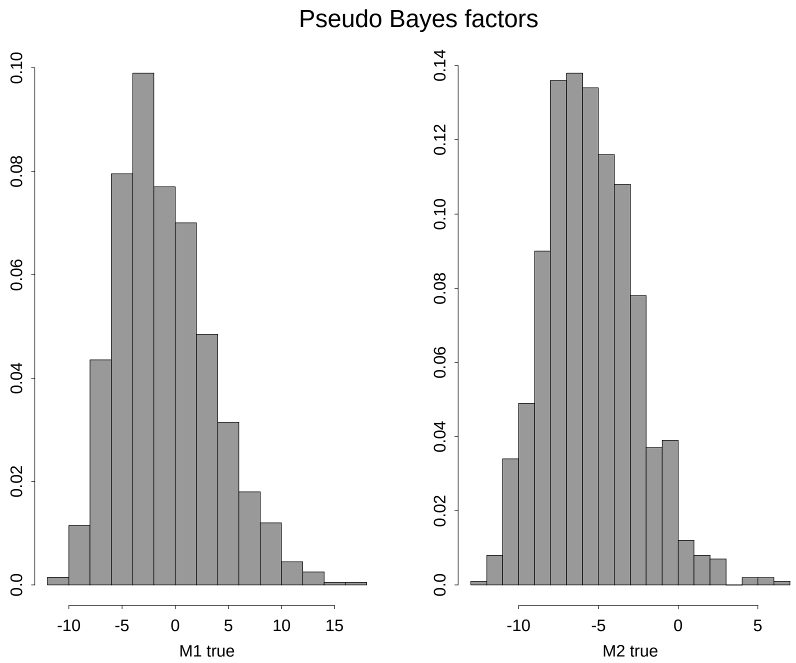

Figure 4: Distributions of $\ln(PsBF)$ under M_1 and M_2 for UNOS data

simple identity: A theoretical exploration. *Statistica Sinica* **6**, 831–860.

Newton, M. A. & Raftery, A. E. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap (with discussion). *Journal of the Royal Statistical Society, Series B, Methodological* **56**, 3–26.

O'Hagan, A. (1995). Fractional Bayes factors for model comparison (with discussion). *Journal of the Royal Statistical Society, Series B* **57**, 99–118.

Sellke, T., Bayarri, M. J., & Berger, J. O. (1999). Calibration of p-values for testing precise null hypotheses. Technical report, Duke University.

Smith, A. F. M. & Spiegelhalter, D. J. (1980). Bayes factors and choice criteria for linear models. *Journal of the Royal Statistical Society, Series B* **42**, 213–220.

Spiegelhalter, D. J. & Smith, A. F. M. (1982). Bayes factors for linear and loglinear models with vague prior information. *Journal of the Royal Statistical Society, Series B*

44, 377–387.

STONE, M. (1970). Necessary and sufficient conditions for convergence in probability to invariant posterior distributions. *The Annals of Mathematical Statistics* **41**, 1349–1353.

TIERNEY, L. & KADANE, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association* **81**, 82–86.

VERDINELLI, I. & WASSERMAN, L. (1995). Computing Bayes factors using a generalization of the Savage-Dickey density ratio. *Journal of the American Statistical Association* **90**, 614–618.