The Canadian Journal of Statistics
Vol. 45, No. 3, 2017, Pages 290–309
La revue canadienne de statistique

Bayes factor biases for non-nested models and corrections

Edward SUSKO*

Department of Mathematics and Statistics, Dalhousie University, Halifax, NS, Canada

Key words and phrases: Bayes factor; bias; fractional Bayes factor; Laplace approximation; model selection; posterior Bayes factor.

MSC 2010: Primary 62F115; secondary 62A01, 62F05

Abstract: With the advent of simulation-based methods to obtain samples from posteriors and due to increases in computational power, Bayesian methods are increasingly applied to complex problems, sometimes providing the only available methods where likelihood implementations are difficult. As a consequence a large body of research in science and social science increasingly utilizes Bayesian tools, often applying them with default settings. A fundamental problem of interest is model selection, and Bayes factors provide a natural approach to Bayesian model selection. Using Laplace approximations and illustrative examples we demonstrate that Bayes factors can have strong biases toward particular models even in non-nested settings with the same number of parameters. Several easily implemented corrections are shown to provide effective cross-checks to default Bayes Factors. *The Canadian Journal of Statistics* 45: 290–309; 2017

Résumé: Grâce aux méthodes de simulation qui permettent d'obtenir des échantillons suivant la loi a posteriori et à la progression rapide de la capacité de calcul, les méthodes bayésiennes sont utilisées de plus en plus fréquemment pour résoudre des problèmes complexes. Elles constituent parfois la seule approche possible lorsque le calcul de la vraisemblance est difficile. Par conséquent, de nombreux travaux de recherche en sciences naturelles et sociales utilisent des méthodes bayésiennes, souvent en conservant la valeur par défaut des paramètres. La sélection de modèle est un problème fondamental d'intérêt et le facteur de Bayes offre une approche naturelle pour le résoudre. À l'aide d'une approximation de Laplace et d'exemples éloquents, l'auteur démontre que le facteur de Bayes peut comporter de forts biais envers certains modèles, même lorsque ceux-ci ne sont pas imbriqués et comportent le même nombre de paramètres. L'auteur montre que plusieurs correctifs faciles à mettre en place offrent une solution efficace aux problèmes du facteur de Bayes par défaut. *La revue canadienne de statistique* 45: 290–309; 2017

1. INTRODUCTION

Model selection is a component of many statistical applications. Within the Bayesian framework, given a collection of models, posterior probabilities of those models are the natural measurements of relative support. For pairwise comparisons the closely associated Bayes factors are frequently used in place of posteriors (Kass & Raftery, 1995; Kadane & Lazar, 2004). Bayes factors are the posterior odds of one model against the other, calculated with equal prior probability for each of the two models.

Bayes factors arise in the Bayesian form of hypothesis testing developed by Jeffreys (1961). Much like in frequentist hypothesis testing it is common that one model is a special case of the other. This setting is unusual in at least two regards. First, Bayes factors for the simpler model can be anomalously large in the presence of disperse priors (Lindley, 1957). The biases considered in

* *Author to whom correspondence may be addressed.*
E-mail: edward.susko@gmail.com

this article are distinct from Lindley's paradox. Second, as one model is nested within another, checking whether credible intervals for more complex models contain parameters consistent with the simpler model provides an alternative means for assessing whether the simpler model is plausible.

For non-nested models, which are the focus of this article, credible regions for parameters no longer provide a means for assessing model plausibility and Bayes factors or posterior probabilities of models are the only direct way of comparing support for the models. We use Laplace approximations and a sequence of examples involving non-nested models to illustrate how and why Bayes factors can lead to bias. We assume that data is generated from a probability distribution corresponding to one of the models considered, which we refer to as the correct model. However, as correct models will always be favoured with sufficient separation between models, the device used to illustrate bias is to consider model spaces that intersect and generate data with parameters for which both models are correct. The device is a proxy for the more general setting of interest where models are similar enough to make discrimination difficult. In the setting of intersecting models we seek situations where Bayes factors are large for one model much more frequently than for the other model even though, in the generating model, both models can be considered correct. Due to continuity of probabilities over parameter spaces the implication is that there will be regions of model space where one model is correct, the other is not and yet Bayes factors favour the incorrect model. This constitutes our definition of model selection bias, a definition consistent with previous extensions of the concept of bias considered in Neyman & Pearson (1936) and Lehman (1951).

In all of the examples considered here models and priors are chosen so that integrals can be explicitly calculated, ensuring that results are not be due to integral approximation. Much effort has been devoted to the construction of priors that are objective (Kass & Wasserman, 1996; Berger, Bernardo, & Sun, 2009; Ghosh, 2011) the goal here is to illustrate difficulties with reasonable choices often made for convenience.

Illuminating the potential for bias is a main goal we also consider several bias corrections. The emphasis is on simple corrections with the two main ones being the posterior Bayes factor of Aitkin (1991) and the fractional Bayes factor of O'Hagan (1995). The corrections can be approximated by averaging exponentiated likelihoods coming from posterior samples of parameters. They thus provide a way of correcting bias without altering existing Bayesian software, without having access to corresponding maximum likelihood implementations and without needing to determine objective priors specific to the problem at hand. In an era of increasingly sophisticated Bayesian implementations applied to complex models the corrections provide a way for end-users to cross-check model selection results through the simple, intuitively appealing strategy of comparing average likelihoods across models.

It should be emphasized that the analysis below is a frequentist study of Bayesian procedures. Because implementations of Bayesian methods are so prevalent, frequentist properties are increasingly of interest to end-users of Bayesian software and to the consumers of the results of Bayesian statistical analyses.

2. THEORY AND METHODS

For all examples considered, data $\boldsymbol{x} = [x_1, \ldots, x_n]^\top$ are assumed independently distributed from densities $p_i(x_i|\theta)$ having unknown parameter θ and leading to a joint density $p(\boldsymbol{x}|\theta) = \prod_i p_i(x_i|\theta)$. The prior, $\pi(\theta)$, is proper with derivatives of all orders and positive density on all possible parameter values. The posterior for θ is then $p(\theta|\boldsymbol{x}) = p(\boldsymbol{x}|\theta)\pi(\theta)/p(\boldsymbol{x})$, where the marginal density, $p(\boldsymbol{x}) = \int p(\boldsymbol{x}|\theta)\pi(\theta)\, d\theta$, is the crucial quantity for Bayes factor calculation.

For model selection it becomes necessary to indicate dependence of $p(\boldsymbol{x}|\theta, m)$, $\pi(\theta|m)$ and $p(\boldsymbol{x}|m)$ on the model m and allow a prior, α_m. The natural measure of support is then the

posterior for model m, $p(m|\boldsymbol{x}) = p(\boldsymbol{x}|m)\alpha_m / p(\boldsymbol{x})$, where now $p(\boldsymbol{x}) = \sum_m p(\boldsymbol{x}|m)\alpha_m$. To compare two models the relative support for Models 1–2 is then the ratio of posteriors, $BF_{12} = p(\boldsymbol{x}|1)\alpha_1 / p(\boldsymbol{x}|2)\alpha_2$. For pairwise comparisons BF_{12} has the advantage that only the ratio α_1/α_2 of priors is needed and the full set of models need not be specified. Setting $\alpha_1 = \alpha_2$ gives BF_{12} as the Bayes factor in which case the posterior for Model 1 is $PF_{12} = BF_{12}/(1 + BF_{12})$. Whether or not $\alpha_1 = \alpha_2$, PF_{12} is a monotone increasing transformation of BF_{12} with the attractive properties of being in [0, 1] and having $PF_{21} = 1 - PF_{12}$. We use it throughout as an alternative but equivalent measure of model support.

We consider model selection allowing for the possibility that for some data sets no model is definitively selected. For Bayes factors $BF_{12} > 1$ is the criteria for Model 1 selection in the conventional sense when a model is always selected. But $BF_{12} > t$, for $t > 1$ is an alternative, more conservative, criteria that leads to no selection when $1/t < BF_{12} < t$.

2.1. A Definition of Bias for Intersecting Models

A desirable property of any model selection procedure is that the probability of selecting the true model is larger than the probability of selecting any other particular model. Model selection procedures satisfying this property are those that we define as "unbiased." The existence of bias depends on the particular parameters in the generating model. It is possible that a model selection procedure will be biased for some parameter settings and unbiased for others. In the two model case when Model 2 is correct and the model selection procedure is biased for the setting considered we say that the model selection procedure is biased toward Model 1. We adopt the (arbitrary) convention throughout that the true model is Model 2 so that bias, if it occurs, is always toward Model 1.

For parameter settings where model separation is large, reasonable methods will heavily favour the correct model, and consequently be unbiased. Thus meaningful investigations of model selection bias (or model selection performance of any sort) need to focus on models that are not well separated. As an idealized framework for models that are not well separated we suppose that the models intersect: there exist θ_{10} and θ_{20} such that $p(\boldsymbol{x}|\theta_{10}, 1) = p(\boldsymbol{x}|\theta_{20}, 2)$. We also assume that the densities are continuous functions of their parameters. This setting provides a means for investigating regions of parameter space where bias occurs.

"Bias Condition": The Bayes factor is biased toward Model 1 when the true model is Model 2 with true parameter $\theta_2 \approx \theta_{20}$ if, for any $t \geq 1$,

$$P(BF_{12} > t) > P(BF_{21} > t) \tag{1}$$

where probabilities are calculated under $p(\boldsymbol{x}|\theta_{20}, 2)$; equivalently $p(\boldsymbol{x}|\theta_{10}, 1)$.

In other words (1) states that Model 1 is more likely to be declared correct regardless of what threshold one uses for declaring a model correct, even when both models are correct. That (1) is sufficient for bias in a neighbourhood of θ_{20}, is a consequence of the continuity of $p(\boldsymbol{x}|\theta_m, j)$ as a function of θ_m: $P(BF_{12} > t; 2, \theta_2) - P(BF_{21} > t; 2, \theta_2)$ converges to $P(BF_{12} > t) - P(BF_{21} > t)$ as $\theta_2 \to \theta_{20}$.

Bias as defined here differs from the usual definition where $\hat{\theta}$ is unbiased when $E_\theta[\hat{\theta}] = \theta$ and otherwise has bias $E_\theta[\hat{\theta}] - \theta$. That definition is natural for quantities of interest that vary continuously but not when the quantity of interest is discrete and unordered as in model selection. Neyman & Pearson (1936) extended the concept of bias to confidence sets and hypothesis tests. A test is unbiased if power under the alternative hypothesis is always at least as large as the type I error rate. Equivalently, using the duality between testing and confidence sets, a confidence set is unbiased if it contains the true parameter with probability at least as large as any other fixed parameter.

This latter definition applies to model selection when confidence sets of models are considered in place of confidence sets of parameters. In a Bayesian setting the analogue of a confidence set is a credible set. Application of the Neyman–Pearson definition is then that the true model is more likely to be contained in the credible set than any other fixed model. If there are only two models, Model 1 is contained in a $(1-\alpha)\times 100\%$ credible set when $BF_{12} > t$ where $t = (1-\alpha)/\alpha$. Similarly Model 2 is in the set if $BF_{21} > t$. Thus the Neyman–Pearson definition requires

$$P(BF_{12} > t; 2, \theta_2) \leq P(BF_{21} > t; 2, \theta_2) \text{ and } P(BF_{21} > t; 1, \theta_1) \leq P(BF_{12} > t; 1, \theta_1). \quad (2)$$

Letting $\theta_1 \to \theta_{10}$ and $\theta_2 \to \theta_{20}$, where θ_{10} and θ_{20} are parameter values at which the two models intersect, (2) gives the requirement that $P(BF_{12} > t) = P(BF_{21} > t)$ when probabilities are calculated under Model 2 and θ_{20}; equivalently Model 1 and θ_{10}. Thus model selection is biased according to Neyman–Pearson if the bias condition (1) is satisfied.

Lehman (1951) considered a decision-theoretic extension of the notion of bias. A decision rule $\delta(\boldsymbol{x})$ is unbiased if the expected loss, $E_\theta[L(\theta', \delta(\boldsymbol{X}))]$, when considered as a function of θ', is minimized by $\theta' = \theta$. For model selection, the decision rule, $\delta(\boldsymbol{X})$, indicates which model has been selected. A natural loss function $L(m', \delta(\boldsymbol{X}))$ is 0 when $\delta(\boldsymbol{x}) = m'$, and 1 otherwise. The expected loss is then $1 - P(\delta(\boldsymbol{X}) = m'; m, \theta_m)$. Consider the decision rule to select a model when its $BF > t$ for some $t > 1$; $t > 1$ allows for the possibility that no decision is made. Then for Bayes factors the Lehman definition of bias gives (2). Letting $\theta_m \to \theta_{m0}$ model selection is biased according to Lehman if the bias condition (1) is satisfied.

One way in which the bias occurs is when Model 1 is nested within Model 2. In that case, when both models are correct, $BF_{12} \to \infty$ as $n \to \infty$ (Schwarz, 1978). Here bias is usually accepted as an attractive feature of Bayesian methods due to Model 1 being more parsimonious; there is some debate as to whether the strength of the bias toward the complex model is too strong (Burnham & Anderson, 2013). However when the numbers of parameters in the two models are the same, which is the focus of this article, parsimony arguments no longer apply, and bias toward Model 1 is undesirable.

A different bias definition was given in Berger & Pericchi (2001) who pointed out that default Bayes procedures, like their intrinsic Bayes factors (Berger & Pericchi, 1996), the Bayesian information criteria (BIC) of Schwarz (1978) and the fractional Bayes factor (FBF) of O'Hagan (1995) are often approximately the same as Bayes factors with conventional priors which, however, do not integrate to 1. Bias in their definition is the ratio of the integrals of the intrinsic priors. Because we assume proper priors none of the examples considered here would be considered bias according to Berger & Pericchi (2001).

2.2. Laplace Approximations and Bias

The Laplace approximation (Tierney & Kadane, 1986) provides insight about bias and why the corrections that we consider are expected to work more generally. We assume the regularity conditions (Kass, Tierney, & Kadane, 1990) for the Laplace approximation. In the preceding discussion the true parameter was θ_{20}, in which case both Models 1 and 2 are correct. To more generally allow for the possibility that Model 2 is correct but Model 1 is not we denote the true parameter as θ_{2n}; $\theta_{2n} = \theta_{20}$ is a special case. Similarly as in Kass & Vaidyanathan (1992) and in the theory for locally most powerful tests, the true parameter is allowed to depend on n and is assumed to satisfy that $\theta_{2n} = \theta_{20} + O(n^{-1/2})$. With greater model separation than this, reasonable methods will heavily favour the correct model (Kass & Vaidyanathan, 1992). In other words the set of θ_2 for which model selection probabilities are non-trivial (not extremely close to 0 and 1) changes with n and is roughly defined through the condition that $\theta_{2n} = \theta_{20} + O(n^{-1/2})$. For this set of true generating parameters Laplace approximations are applicable. Thus bias correction methods motivated by Laplace approximations are expected to work at eliminating bias for those

model parameters that give non-trivial model selection probabilities. Outside of these parameter settings, bias-corrected methods, Bayes factors, and other model selection procedures give large probabilities of selecting the correct model anyways.

Let d_m be the dimension of θ_m and let $\hat{\theta}_m$ be the maximizer of the likelihood, $L_m(\theta) = p(\boldsymbol{x}|\theta, m)$. Denote the log likelihood $l_m(\theta) = \log L_m(\theta)$ and let $I_m(\theta) = E[-l_m^{(2)}(\theta)/n|\theta, m]$ be the expected information matrix. Then the Laplace approximation (Tierney & Kadane, 1986; Kass, Tierney, & Kadane, 1990) gives that for any fixed $b > 0$ and $m = 1, 2$,

$$\int [p(\boldsymbol{x}|\theta, m)]^b \pi(\theta|m)\, d\theta = \left(\frac{2\pi}{bn}\right)^{d_m/2} \pi(\hat{\theta}_m|m)|I_m(\hat{\theta}_m)|^{-1/2}[L_m(\hat{\theta}_m)]^b \times [1 + O_p(n^{-1/2})]. \tag{3}$$

The reason for allowing $b \neq 1$ will become clear presently. Smaller errors in the approximation are achievable by replacing $\hat{\theta}_m$ with other values (Tierney & Kadane, 1986) but errors with $\hat{\theta}_m$ are small enough that Laplace approximations can be used to show that bias is generally expected and to motivate relatively simple corrections.

Suppose that Model 2 is correct and has true parameter θ_{2n} satisfying that $\theta_{2n} = \theta_{20} + O(n^{-1/2})$. Then (3) gives

$$\frac{\int [p(\boldsymbol{x}|\theta, 1)]^b \pi(\theta|1)\, d\theta}{\int [p(\boldsymbol{x}|\theta, 2)]^b \pi(\theta|2)\, d\theta} = \left(\frac{2\pi}{bn}\right)^{(d_1-d_2)/2} \frac{\pi(\hat{\theta}_1|1)}{\pi(\hat{\theta}_2|2)} \frac{|I_1(\hat{\theta}_1)|^{-1/2}}{|I_2(\hat{\theta}_2)|^{-1/2}} \left[\frac{L_1(\hat{\theta}_1)}{L_2(\hat{\theta}_2)}\right]^b + O_p(n^{-1/2}). \tag{4}$$

The Bayes factor, BF_{12}, is approximated by (4) with $b = 1$.

If $d_1 \neq d_2$, assuming without loss of generality that $d_1 < d_2$, $O_p(1)$ contributions to (4) are obtained from terms involving the priors, involving $I_m(\hat{\theta}_m)$ and, as $\theta_{2n} = \theta_{20} + O(n^{-1/2})$, even from the term $L_1(\hat{\theta}_1)/L_2(\hat{\theta}_2)$. The ratio (4) is then $n^{(d_2-d_1)/2}$ times an $O_p(1)$ quantity and will increase without bound as sample size gets large.

When $d_1 = d_2$, which is the focus of this article, the likelihood ratio term in (4) will contribute to variability but is not expected to be an appreciable source of bias. The main potential sources of bias in this case, and an additional source in the case $d_1 \neq d_2$, comes from the product of the prior ratio and the information ratio,

$$b(\hat{\theta}_1, \hat{\theta}_2) = \frac{\pi(\hat{\theta}_1|1)}{\pi(\hat{\theta}_2|2)} \frac{|I_1(\hat{\theta}_1)|^{-1/2}}{|I_2(\hat{\theta}_2)|^{-1/2}}. \tag{5}$$

The following result is proved in the appendix.

Theorem 1. *Assume the regularity conditions of Kass, Tierney, & Kadane (1990), that* $d := d_1 = d_2$ *and that* $\theta_{2n} = \theta_{20}$. *Then*

$$P(BF_{12} > t) \to P(Vb(\theta_{10}, \theta_{20}) > t) \text{ as } n \to \infty,$$

where $2\log(V)$ *has a symmetric distribution around 0 which corresponds to a distribution of a difference of two correlated* χ^2_d *random variables. For any* $t > 0$,

$$P(BF_{12} > tb(\theta_{10}, \theta_{20})^2) - P(BF_{21} > t) \to 0 \text{ as } n \to \infty.$$

If $b(\theta_{10}, \theta_{20}) > 1$, then $P(BF_{12} > t) > P(BF_{12} > tb(\theta_{10}, \theta_{20})^2)$ for all large n, so

$$\lim_{n\to\infty} P(BF_{12} > t) - P(BF_{21} > t) > \lim_{n\to\infty} P(BF_{12} > tb(\theta_{10}, \theta_{20})^2) - P(BF_{21} > t) = 0,$$

implying bias. Thus the main potential sources of bias in this case, and an additional source in the case $d_1 \neq d_2$, comes from $b(\theta_{10}, \theta_{20}) \approx b(\hat{\theta}_1, \hat{\theta}_2)$. In addition because of the symmetry of the distribution of $\log(V)$, asymptotically, the median BF_{12} value is $b(\theta_{10}, \theta_{20})$.

The potential for bias is a consequence of the product of the prior ratio $\pi(\hat{\theta}_1|1)/\pi(\hat{\theta}_2)$ and the information ratio $|I_1(\hat{\theta}_1)|^{-1/2}/|I_2(\hat{\theta}_2)|^{-1/2}$. Because the choice of a prior will vary across practitioners, the importance of the information ratio in causing bias is emphasized in the examples below. However the contribution of the information ratio to the product is not well-defined when one allows for different parameterizations. In all of the examples below we consider parameterizations that are natural for some data settings.

2.3. Bias Corrections

For a given parameterization the bias from the information ratio is an inherent bias not directly due to prior specification. It leads to the first correction which specifies a model prior $\alpha_m \propto |I_m(\hat{\theta}_m)|^{1/2}$ and uses the posterior odds ratio, $BF_{12}\alpha_1/\alpha_2$ in place of BF_{12}. A cancellation occurs upon multiplying (4) by α_1/α_2, causing the $|I_m(\hat{\theta}_m)|^{-1/2}$ terms to vanish. As model priors require calculation of $I_m(\hat{\theta}_m)$ or an observed information approximation, the correction may be difficult to implement in complex settings or when end-user access is only to the output of Bayesian software. We consider the model prior correction partly because it only corrects the bias due to the information ratio. When it shows good results relative to BF_{12} it implies that bias is due to the information ratio term.

The posterior Bayes factor (PBF) of Aitkin (1991) replaces BF_{12} with

$$PBF = \frac{\int [p(\boldsymbol{x}|\theta, 1)]^b \pi(\theta|1)\, d\theta}{\int p(\boldsymbol{x}|\theta, 1)\pi(\theta|1)\, d\theta} \times \frac{\int p(\boldsymbol{x}|\theta, 2)\pi(\theta|2)\, d\theta}{\int [p(\boldsymbol{x}|\theta, 2)]^b \pi(\theta|2)\, d\theta}, \tag{6}$$

where $b = 2$ in Aitkin (1991); we allow a slight generalization to $b > 1$ in order to make a connection to the fractional Bayes factor (FBF) of O'Hagan (1995). Approximating terms in (6) by (4) gives that

$$PBF = \left(\frac{1}{b}\right)^{(d_1-d_2)/2} \left[\frac{L_1(\hat{\theta}_1)}{L_2(\hat{\theta}_2)}\right]^{b-1} + O_p(n^{-1/2}). \tag{7}$$

Due to the ratios considered the bias factor $b(\hat{\theta}_1, \hat{\theta}_2)$ in (4) cancelled.

An attractive feature of PBF is that it is simply approximated by an intuitively appealing measure of support. In the original form of Aitkin (1991),

$$\int p(\boldsymbol{x}|\theta, m)^b \pi(\theta|m)\, d\theta \Big/ \int p(\boldsymbol{x}|\theta, m)\pi(\theta|m)\, d\theta$$

was expressed as $E([p(\boldsymbol{x}|\theta, m)]^{b-1}|\boldsymbol{x}, m)$. A wide variety of methods are available to simulate $\theta_{m1}, \ldots, \theta_{mB}$ from $p(\theta|\boldsymbol{x}, m)$ (Gilks, Richardson, & Speigelhalter, 1996). When, as is frequently the case, the Bayesian implementation outputs log likelihoods for each of the simulated θ_j, PBF can be approximated by the ratio of the average likelihoods

$$B^{-1}\sum_{k=1}^{B} L_1(\theta_{1k})^{b-1} \Big/ B^{-1}\sum_{k=1}^{B} L_2(\theta_{2k})^{b-1}. \tag{8}$$

Even when the simulation output, $(\theta_1, m_1), \ldots, (\theta_B, m_B)$, is from $p(\theta, m|\boldsymbol{x})$, the ratio

$$\sum_{k|m_k=1} L_1(\theta_k)^{b-1} \Big/ \sum_{k|m_k=2} L_2(\theta_k)^{b-1} \tag{9}$$

provides an approximation.

Closely associated with the PBF is the fractional Bayes factor (FBF) of O'Hagan (1995),

$$FBF = \frac{\int p(\boldsymbol{x}|\theta, 1)\pi(\theta|1)\, d\theta}{\int [p(\boldsymbol{x}|\theta, 1)]^b \pi(\theta|1)\, d\theta} \times \frac{\int [p(\boldsymbol{x}|\theta, 2)]^b \pi(\theta|2)\, d\theta}{\int p(\boldsymbol{x}|\theta, 2)\pi(\theta|2)\, d\theta} \tag{10}$$

where $b < 1$. Approximating terms in (10) by (4),

$$FBF = b^{(d_1-d_2)/2} \left[\frac{L_1(\hat{\theta}_1)}{L_2(\hat{\theta}_2)}\right]^{1-b} + O_p(n^{-1/2}).$$

Thus, when $d_1 = d_2$, FBF with $b = 1 - \delta$ agrees with PBF using $b = 1 + \delta$ up to $O_p(n^{-1/2})$.

A criticism that has been raised about PBF is that it is not Bayesian. Suppose now that d_1 is not necessarily the same as d_2 and consider a simple transformation of PBF

$$TPBF = \{b^{(d_1-d_2)/2} PBF\}^{1/(b-1)} n^{(d_2-d_1)/2}$$

when $b = 2$ and $d_1 = d_2$ this is just PBF. For $b \leq 2$, from (7), up to $O_p(n^{-1/2})$, $TPBF$ is

$$\exp(S) = \frac{L_1(\hat{\theta}_1)}{L_2(\hat{\theta}_2)} n^{(d_2-d_1)/2} \tag{11}$$

which is the BIC approximation to the Bayes factor given by Schwarz (1978). Thus TPBF is Bayesian in the same sense as BIC. Kass & Wasserman (1995) argued that $\exp(S)$ is generally a crude approximation to Bayes factors but that, for null orthogonal parameters and nested models, $\exp(S)$ is approximately the Bayes factor corresponding to a normal prior distribution. A similar argument applies here. Suppose the prior for $\theta_m|m$ is $N(\theta_{m0}, I(\theta_{m0})^{-1/2})$, $m = 1, 2$. Then, as $\hat{\theta}_m = \theta_{m0} + O_p(n^{-1/2}), \pi(\hat{\theta}_m|m) = (2\pi)^{-d_m/2}|I_m(\theta_{m0})|^{1/2} + O_p(n^{-1/2})$. Substituting this approximation to $\pi(\hat{\theta}_m|m)$ in (4) with $b = 1$ gives $\exp(S)$ up to $O_p(n^{-1/2})$, which is also TPBF up to $O_p(n^{-1/2})$. As these priors cannot be used directly without knowledge of the true θ_{m0}, they exist in the setting of concern where models are close to each other, and they are reasonable. Thus a motivation for $TPBF$ is that it approximates Bayes factors corresponding to $N(\theta_{m0}, I(\theta_{m0})^{-1/2})$ priors without knowledge of θ_{m0}.

3. EXAMPLES

Simplified examples of more general settings where Bayes factors are expected to be biased toward Model 1 are considered. A summary of the examples is given below.

3.1 For two structurally similar models of the same dimension, Bayes factors are biased toward the model with greater variance in parameter estimation.

3.2 Homogeneity of outcome probabilities is often of interest (e.g., a disease case is as likely to be from region 1 as from region 2) which leads to the comparison of models that com-

bine categories. Bayes factors are biased toward models that combine sparsely represented categories.

3.3 In regression models Bayes factors tend to be biased toward models with highly correlated predictors.

3.4 Parametric relationships are sometimes not compatible across groups (e.g., different regression relationships for different genders) leading to models that stratify according to the group membership. Bayes factors tend to be biased toward models that have more unequal strata.

3.5 Based on experience with nested models, it is expected that when both models are correct, Bayes factors will favour the model with fewer parameters. However, even with moderate sample sizes, models with greater variability in estimation can counteract the tendency to favour lower dimensional models.

3.1. Variation in Parameter Estimation Variability

In the first example the data consists of independent and identically distributed $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$, where the X_i have a $N(\mu_x, 1)$ distribution and the Y_i a $N(\mu_y, \sigma_y^2)$ distribution with $\sigma_y^2 > 1$ known. In Model 1 $\mu_x = 0$ and μ_y potentially differs from 0, whereas in Model 2 $\mu_y = 0$ and μ_x potentially differs from 0. This is a simple illustrative example of a situation where maximum likelihood estimation under Model 1 (estimation of μ_y) is more variable than under Model 2 (estimation of μ_x), which turns out to be a major determining factor as to whether bias can be expected. The models might arise, for instance, in a comparison of treatments x and y using paired differences. In such a setting the X_i and Y_i would represent differences in measurements (after–before) for independent samples of individuals. Under Model 1 treatment y has some effect but x does not and under Model 2, treatment x has some effect but y does not. We assume a $N(0, 1)$ prior for the unknown μ_y in Model 1, and the same prior for the unknown μ_x in Model 2.

The bias factors in (5) are calculated as $\pi(\hat{\theta}_1|1)/\pi(\hat{\theta}_2|2) = 1$ and $|I_1(\hat{\theta}_1)|^{-1/2}/|I_2(\hat{\theta}_2)|^{-1/2} = \sigma_y$. If $\sigma_y \gg 1$, Model 1 is expected to be favoured. This is confirmed in Table 1 which gives the probabilities with which $PF_{12} = BF_{12}/(1 + BF_{12})$ and PF_{21} are greater than thresholds 0.5, 0.6, 0.75, and 0.9. Tabulated probabilities correspond to $\mu_x = \mu_y = 0$ and $n = 50$. As in all probability approximations throughout the article, $B = 10{,}000$ data sets were simulated. The standard deviation of the approximation to a probability, P, is then $\sqrt{P(1-P)/1{,}000} \leq \sqrt{0.5(1-0.5)/10{,}000} = 0.005$.

Bias toward Model 1 occurs when the percentages of $PF_{12} > t$ (first number) are larger than $PF_{21} > t$ (second number). As expected BF is biased toward Model 1. Posterior Bayes factors and FBF (with $b > 0.01$) are effective in correcting the bias. An exception occurs in the extreme case that $\sigma_y = 5$. In this case the FBF correction with $b = 0.01$ is not effective and bias toward Model 1 remains. In theory, with large samples, results from FBF should be comparable to PBF with $b = 1.99 \approx 2$, but this is not the case. By contrast, when $\sigma_y = 5$, PBF shows little evidence of bias. The problem for theory is that the errors in the Laplace approximations to both PBF and FBF decrease with n largely through nb. As nb tends to be large for PBF but small for FBF, approximations are poor for FBF.

Model prior corrections, $\alpha_m \propto |I_m(\hat{\theta}_m)|^{1/2}$, correct for bias when $\sigma_y < 5$. As model prior corrections do not involve priors for parameters and only correct for bias due to the information ratio, the results with $\sigma_y < 5$ confirm that the bias of BF is due to the information ratio. Using model priors over-corrects when $\sigma_y = 5$ where the chance of selecting Model 2 is actually larger than the chance of selecting Model 1. This is due to a small sample bias. When $n = 500$ the corresponding entries are 45/55, 21/25, 11/14, 6/7, and 2/3; that is, the model-prior corrected proportion of $PF_{12} > t$ is comparable to $PF_{21} > t$.

The example illustrates the first type of setting where bias is generally expected. As $|I_m(\hat{\theta}_m)|^{-1/2}/\sqrt{n}$ is the standard error of the maximum likelihood estimator for Model m,

TABLE 1: The percentages of times $BF_{12}/(1+BF_{12})$ (first number) or $BF_{21}/(1+BF_{21})$ (second number) was larger than 0.5, 0.6, 0.75, and 0.9 when Models 1 and 2 are both correct in Example 1.

	$\sigma_y = 1.5$				$\sigma_y = 2.0$				$\sigma_y = 5.0$			
	0.50	0.60	0.75	0.90	0.50	0.60	0.75	0.90	0.50	0.60	0.70	0.90
BF	76/24	47/14	15/5	3/1	83/17	70/10	21/4	5/1	94/6	89/4	70/2	8/0
Model prior	47/53	22/24	8/9	2/2	44/56	21/25	7/10	2/3	25/75	11/35	3/13	0/3
FBF $b = 0.01$	72/28	38/16	12/6	3/2	80/20	61/12	16/5	4/1	92/8	86/5	59/2	5/1
FBF $b = 0.5$	53/47	14/13	2/2	0/0	56/44	14/13	2/2	0/0	69/31	16/10	2/2	0/0
PBF $b = 1.5$	51/49	13/13	2/2	0/0	52/48	13/13	2/2	0/0	61/39	14/12	2/2	0/0
PBF $b = 2$	51/49	24/24	9/9	2/2	52/48	24/23	9/9	2/3	59/41	26/21	9/8	2/2

then if the estimates for Model 1 tend to be more variable than estimates for Model 2, $|I_1(\hat{\theta}_1)|^{-1/2}/|I_2(\hat{\theta}_2)|^{-1/2}$ will be large and bias in favour of Model 1 is expected.

The satisfaction of the condition (1) for bias exhibited in Table 1 implies that bias will result when Model 2 is correct and Model 1 is not. This is seen in the first row of Figure 1 which give the probabilities of correctly estimating Model 2 as μ_x increases away from 0; all other parameter settings remain the same as for Table 1. For unbiased model selection, the y-values for the curves should always remain at least as large as 1/2. With the exception of FBF and $b = 0.01$, the corrected approaches involving PBF and FBF all show only slight biases. BF (and FBF $b = 0.01$) show substantial bias that increases as the information ratio, or equivalently σ_y, increases. One sees the importance of the condition (1) in determining bias. Model selection favours Model 2 as μ_x increases away from 0 so that it is the initial y-axis value that determines whether there will be a region of bias.

3.2. Homogeneity of Multinomial Probabilities

In the second example X_1, X_2, X_3, and X_4 are multinomial with size parameter n and multinomial probabilities p_1, p_2, p_3, and p_4. Model 1 postulates that $p_1 = p_2$ whereas Model 2 that $p_3 = p_4$. Such a model comparison might arise, for instance, when the p_j represent the probability of a disease case from a region j. A common question of interest in such cases is whether and which regions have comparable disease rates. This leads to models that combine separate regions into a single region. An example comparison of interest is the one above where under Model 1 those probabilities do not vary across regions 1 and 2, whereas under Model 2 they do not vary over regions 3 and 4.

We assume a Dirichlet prior, $D(\alpha_1, \alpha_2, \alpha_3, \alpha_4)$ for p_1, p_2, p_3. Under Model 1 this gives a $D(\alpha_1 + \alpha_2, \alpha_3, \alpha_4)$ distribution for the unknown parameters $p_1 + p_2$, p_3, and p_4. Under Model 2 a $D(\alpha_1, \alpha_2, \alpha_3 + \alpha_4)$ prior for p_1, p_2, and $p_3 + p_4$ results. In examples below the α parameters were chosen so that the mean p_i was equal to its true value, p_{i0}, with variance $p_{i0}(1 - p_{i0}) \times 0.9$. As the variance of a beta-distributed random variable with a mean of m is always less than $m(1 - m)$ this gives a disperse prior within the Dirichlet family. The p_{i0} depend on the simulation scenario.

Table 2 gives the probabilities that PF_{12} and PF_{21} exceed thresholds when both models are correct and $n = 500$. As $p_1 = p_2$ and $p_3 = p_4$, the p_i are determined by a single parameter, $\zeta = p_1 + p_2$, through $p_1 = p_2 = \zeta/2$, $p_3 = p_4 = (1 - \zeta)/2$. As $p_1 + p_2$ decreases from 0.5, the bias of BF toward Model 1 increases. FBF with $b = 0.5$ and PBF are very effective at reducing the bias whereas using model priors gives a slight over-correction. As model priors only correct for the information ratio, their effectiveness indicates that the cause of bias is the information ratio.

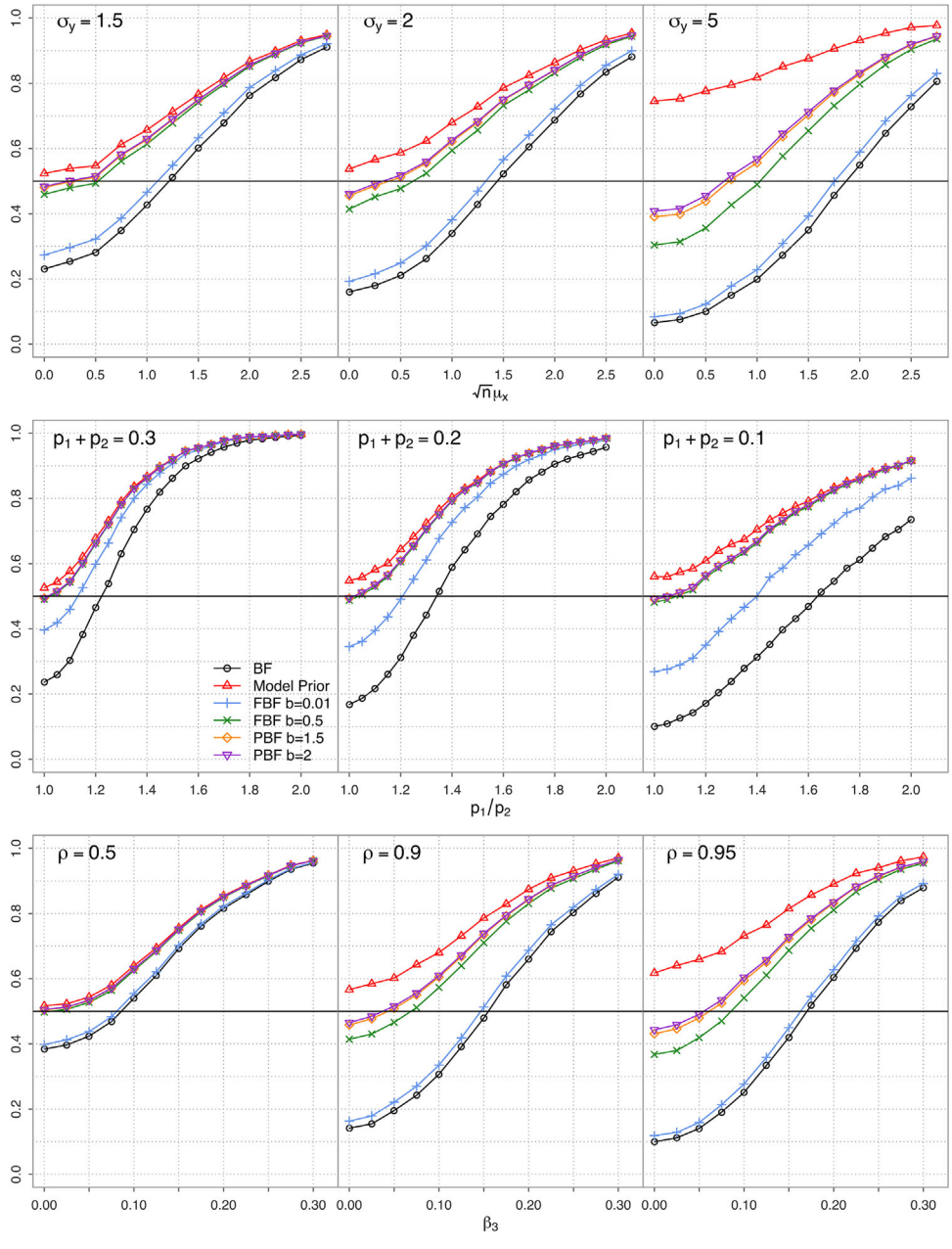

FIGURE 1: For examples 1–3 the frequency with which the correct Model 2 was selected (BF $>$ 1) as parameters increase away from those for which Models 1 and 2 are equivalent.

The second row of Figure 1 gives the probabilities of correctly estimating Model 2 as p_1/p_2 moves away from 1. As with Example 1, the curves illustrate a lack of sensitivity to the choice of b, except for FBF and small b. The ratio $|I_1(\hat{\theta}_1)|^{-1/2}/|I_2(\hat{\theta}_2)|^{-1/2}$ is $[1-(p_1+p_2)]^{1/2}/[p_1+p_2]^{1/2}+O_p(n^{-1/2})$. Consistent with the results of Table 2 and Figure 1 the information ratio, and consequently bias, increase as p_1+p_2 decreases.

The example is a more subtle illustration of a general reason for bias: differences in parameter estimation variability for one model in comparison to another. Aggregating counts for groups

TABLE 2: The percentages of times $BF_{12}/(1+BF_{12})$ (first number) or $BF_{21}/(1+BF_{21})$ (second number) was larger than 0.5, 0.6, 0.75, and 0.9 when Models 1 and 2 are both correct in Example 2.

	$p_1+p_2=0.3$				$p_1+p_2=0.2$				$p_1+p_2=0.1$			
	0.50	0.60	0.75	0.90	0.50	0.60	0.75	0.90	0.50	0.60	0.70	0.90
BF	76/24	49/14	15/6	4/2	83/17	69/10	21/4	5/1	90/10	82/7	43/3	8/1
Model prior	47/53	23/25	9/10	2/3	45/55	21/26	9/10	2/3	45/55	22/28	9/11	2/3
FBF $b=0.01$	61/39	27/21	10/8	2/2	65/35	29/19	11/7	3/2	74/26	42/14	14/6	3/2
FBF $b=0.5$	51/49	13/13	2/2	0/0	51/49	13/14	2/2	0/0	52/48	14/13	2/2	0/0
PBF $b=1.5$	51/49	13/13	2/2	0/0	50/50	13/14	2/2	0/0	51/49	13/13	2/2	0/0
PBF $b=2$	51/49	24/23	9/9	2/2	50/50	23/24	9/9	2/2	51/49	23/23	9/9	2/2

that are sparsely represented gives more variable parameter estimates than aggregating well-represented groups. For instance, when $p_1+p_2=0.1$, the proportions being estimated under Model 1 are $p_1+p_2=0.1$, $p_2=0.45$, and $p_3=0.45$, whereas they are $p_1=0.05$, $p_2=0.05$, and $p_3+p_4=0.90$ under Model 2. The variance of a proportion estimator of p increases from 0 to $1/(4n)$ as p moves from 0 or 1 to $1/2$.

3.3. Correlated Predictors

The third example is a linear regression example involving variable selection. Comparison is between a Model 1 involving covariate vector (x_i, z_i) and Model 2 involving covariate (x_i, w_i). The covariates x_i and w_i are independent of each other, whereas x_i and z_i are correlated.

More formally, in Model 1, $y_i = x_i\beta_1 + z_i\beta_2 + \epsilon_i$ where $\epsilon_1, \ldots, \epsilon_n$ are independent and identically distributed $N(0, \sigma^2)$ variates. Model 2 includes the covariate x_i but replaces z_i with w_i: $y_i = x_i\beta_1 + w_i\beta_3 + \epsilon_i$. In the model generating the data $[x_i, w_i, z_i]^\top$ are jointly normally distributed with marginal zero means and unit variances. The covariate w_i is independent of z_i and x_i but x_i and z_i have correlation ρ. The priors for both models are defined through a conditional $N([1,0]^\top, \sigma^2 I)$ distribution for β given σ and a gamma prior for $1/\sigma^2$ with shape and rate parameters both set to 1/2.

The results in Table 3 correspond to settings where both models are correct. Specifically $n=100$, $\sigma^2=1$, $\beta_1=1$, and $\beta_2=0$. As with previous examples Table 3 indicates bias toward Model 1 and the bias is corrected for by PBF and FBF when $b>0.01$ but under-corrected when $b=0.01$. As model priors correct for information ratio alone and yet over-corrected the BF bias, it is clear that the bias of BF is consequence of the information ratio between the two models, with priors for parameters having a mitigating effect that caused the model priors to over-correct. In Figure 1 β_3 varies away from 0 and only Model 2 is correct. In both Table 3 and Figure 1 the strength of the bias increases as the correlation between x_i and z_i increases.

The example illustrates a general setting where bias toward certain models is expected to arise. Once again the issue is variability of parameter estimation. In this case, however, the problem is that highly correlated predictors give rise to large correlations in parameter estimates.

3.4. Stratification

The fourth example illustrates that bias is expected when comparing stratified models where individuals are more unequally distributed to strata under Model 1 than in Model 2. We begin with a specific example and then argue that the phenomenon is more general.

TABLE 3: The percentages of times $BF_{12}/(1+BF_{12})$ (first number) or $BF_{21}/(1+BF_{21})$ (second number) was larger than 0.5, 0.6, 0.75, and 0.9 when Models 1 and 2 are both correct in Example 3.

	$\rho = 0.5$				$\rho = 0.9$				$\rho = 0.95$			
	0.50	0.60	0.75	0.90	0.50	0.60	0.75	0.90	0.50	0.60	0.70	0.90
BF	62/38	29/20	10/9	3/2	86/14	75/8	26/4	6/1	90/10	83/6	45/3	6/1
Model prior	49/51	23/24	9/10	2/3	43/57	21/25	8/10	2/3	37/63	17/28	6/11	1/3
FBF $b = 0.01$	60/40	27/21	10/9	3/3	84/16	70/9	22/4	5/1	88/12	80/7	33/3	5/1
FBF $b = 0.5$	51/49	13/14	2/3	0/0	59/41	15/12	3/2	0/0	63/37	15/11	3/2	0/0
PBF $b = 1.5$	50/50	13/14	2/3	0/0	54/46	14/13	3/2	0/0	57/43	14/13	3/2	0/0
PBF $b = 2$	50/50	23/24	9/10	2/3	54/46	24/23	10/9	3/2	55/45	25/22	10/9	3/2

In the specific example data is binary, for instance, a disease status indicator. Individuals are stratified according to auxiliary information that differs for the two models. For instance in one model individuals might be stratified according to gender and in the other geographical location. Specifically, under model m, the auxilliary variable, W_m, is either 1 or 0. If, for instance, W_1 indicates gender and W_2 geographical location, then under Model 1 stratification according to gender is necessary and under Model 2 stratification according to geographical location is necessary. For simplicity we assume independence of W_1 and W_2.

The likelihood for the data under Model m is

$$\prod_w p_{s|wm}^{y_{wm}}(1-p_{s|wm})^{n_{wm}-y_{wm}}$$

where $p_{s|wm}$ is the success probability for an individual when $W_m = w$. Here y_{wm} is the number of successes and n_{wm} is the number of individuals with $W_m = w$. The priors for $p_{s|wm}$ were taken as beta with mean equal to the true $p_{s|wm}$ and variances $0.9\,p_{s|wm}(1-p_{s|wm})$. This gives a disperse prior as the variance of a beta-distributed random variable with a mean of m is always less than $m(1-m)$.

The two models are both correct if no stratification is needed: $p_{s|wm} = p_s$. There are three parameters determining distributional properties in this case, p_0, $p_{0|1} = P(W_1 = 0)$ and $p_{0|2} = P(W_2 = 0)$. The information ratio is

$$|I_1(\hat{p}_1)|^{-1/2}/|I_2(\hat{p}_2)|^{-1/2} = \sqrt{p_{0|2}(1-p_{0|2})}/\sqrt{p_{0|1}(1-p_{0|1})} + O_p(n^{-1/2}).$$

As $p(1-p)$ is unimodal and symmetric about its maximizing value $p = 0.5$, bias toward Model 1 is expected when, relative to Model 2, strata are more uneven: $p_{0|1}$ is farther than $p_{0|2}$ from the 0.5.

Table 4 gives the bias results for the example with $n = 100$, $p_0 = 0.5$, $p_{0|2} = 0.5$, and $p_{0|1} = 0.3, 0.2$, or 0.1. As expected the bias toward Model 1 increases as $p_{0|1}$ decreases. Model priors PBF and FBF all provide adequate corrections with the exception of FBF and $b = 0.01$. In contrast to previous examples, however, the FBF bias when $b = 0.01$ is toward Model 2 rather than Model 1.

The example generalizes to other settings involving stratification but the argument for bias is a little different than the one considered in Section 2.2. Consider two models that use the same probability model across strata but allow completely different parameters for different strata. Suppose that the difference between the models is how they assign individuals to strata but, to

TABLE 4: The percentages of times $BF_{12}/(1+BF_{12})$ (first number) or $BF_{21}/(1+BF_{21})$ (second number) was larger than 0.5, 0.6, 0.75, and 0.9 when Models 1 and 2 are both correct in Example 4.

	$p_{0\|1}=0.3$				$p_{0\|1}=0.2$				$p_{0\|1}=0.1$			
	0.50	0.60	0.75	0.90	0.50	0.60	0.75	0.90	0.50	0.60	0.70	0.90
BF	60/40	27/21	10/8	3/2	68/32	35/18	12/7	3/2	80/20	62/11	19/5	4/1
Model prior	50/50	24/24	9/9	3/3	51/49	24/24	9/9	2/3	51/49	25/23	10/9	2/3
FBF $b=0.01$	47/53	23/25	9/9	2/3	43/57	22/26	8/10	2/3	37/63	19/29	8/10	2/3
FBF $b=0.5$	50/50	13/14	2/3	0/0	50/50	13/14	2/3	0/0	49/51	13/13	2/3	0/0
PBF $b=1.5$	50/50	13/14	2/3	0/0	50/50	13/14	2/3	0/0	50/50	13/13	2/3	0/0
PBF $b=2$	50/50	23/24	9/9	2/3	50/50	24/24	9/9	2/3	50/50	24/23	9/9	2/3

keep the parameter dimensions of the models constant, assume the same number of strata within each model. Finally assume enough observations are collected in each stratum to justify separate Laplace approximations to the separate likelihoods for separate strata. Then for model m, having within-stratum data $\boldsymbol{x}_{sm}$ and within-stratum maximum likelihood estimator, $\hat{\theta}_{sm}$, the marginal probability is

$$p(\boldsymbol{x}|m) \approx \prod_s \left(\frac{2\pi}{bn_{sm}}\right)^{d_m/2} p(\boldsymbol{x}_{sm}|\hat{\theta}_{sm}, m)\pi(\hat{\theta}_{sm}|sm)|I_m(\hat{\theta}_{sm})|^{-1/2} \tag{12}$$

where n_{sm} is the number of observations in the sth stratum for model m. When $d := d_1 = d_2$, (12) gives the Bayes factor as

$$BF_{12} \approx \left(\frac{\prod_s n_{s2}/n}{\prod_s n_{s1}/n}\right)^{d/2} \times \frac{\prod_s \pi(\hat{\theta}_{s1}|s1)}{\prod_s \pi(\hat{\theta}_{s2}|s2)} \times \frac{|I_1(\hat{\theta}_{s1})|^{-1/2}}{|I_2(\hat{\theta}_{s2})|^{-1/2}} \times L_R \tag{13}$$

where

$$\log(L_R) = \sum_s \log p(\boldsymbol{x}_{s1}|\hat{\theta}_{s1}, 1) - \sum_s \log p(\boldsymbol{x}_{s2}|\hat{\theta}_{s21}, 2).$$

The two models intersect when stratification is unnecessary: $\theta_{sm} = \theta_0$. In this case, because the models only differ in how they stratify, the I_m contributions in the third factor of (13) are, up to $O_p(n^{-1/2})$ the same for the two models. This is in contrast to previous examples and the arguments in Section 2.2. In any case, as $\hat{\theta}_{sm} \approx \theta_0$, then (13) gives

$$BF_{12} \approx \left(\frac{\prod_s n_{s2}/n}{\prod_s n_{s1}/n}\right)^{d/2} \times \frac{\prod_s \pi(\theta_0|s1)}{\prod_s \pi(\theta_0|s2)} \times L_R. \tag{14}$$

The log likelihood ratio statistics, $2\{\log p(\boldsymbol{x}_{sm}|\hat{\theta}_{sm}, m) - \log p(\boldsymbol{x}_{sm}|\theta_0, m)\}$, within strata, have limiting chi-squared distributions. This implies that the contribution L_R in (14) coming from the likelihoods,

$$\log(L_R) = \sum_s [\log p(\boldsymbol{x}_{s1}|\hat{\theta}_{s1}, 1) - \log p_0(\boldsymbol{x}_{s1})] - \sum_s [\log p(\boldsymbol{x}_{s2}|\hat{\theta}_{s21}, 2) - \log p_0(\boldsymbol{x}_s)]$$

will cause the Bayes factors to vary but should not cause substantial biases; the mean $\log(L_R)$ term is approximately 0. Thus aside, from biases that may be due to priors (which should not occur if priors are comparable across strata), the main potential source of bias from (14) comes from the first term. The bias to Bayes factors expected from this terms is

$$\left(\frac{\prod_s n_{s2}/n}{\prod_s n_{s1}/n}\right)^{d/2} \approx \left(\frac{\prod_s p_{s|2}}{\prod_s p_{s|1}}\right)^{d/2}$$

where $p_{s|m}$ gives the probability of being in stratum s for model m. As $\prod_s p_{s|m}$ is maximal when the p_{sm} are all equal, one can conclude that there will be generally be bias toward Model 1 when comparing stratified models where individuals are more unequally distributed to strata under Model 1 than Model 2.

3.5. Models of Differing Dimension

The final example is a variation of Example 1 that compares models of different dimension. As in Example 1 independent and identically distributed $X_1, \ldots, X_n$ and $Y_1, \ldots Y_n$ are from $N(\mu_x, \sigma_x^2)$ and $N(\mu_y, \sigma_y^2)$ distributions. In Model 1 $\mu_x = 0$, σ_x, and σ_y are fixed at their true values and μ_y has a $N(0, 1)$ prior. In Model 2 $\mu_y = 0$ and σ_y is fixed at its true value but, by contrast with Example 1, σ_x is not fixed. Rather $1/\sigma_x^2$ has a gamma distribution with shape parameter a and rate b. Conditional upon σ_x^2 the prior for μ_x is $N(0, \sigma_x^2)$. For the results reported in Table 5 both Models 1 and 2 are correct with $\mu_x = \mu_y = 0$, $\sigma_x = 1$, and $n = 50$. The gamma parameters were chosen so that $1/\sigma_x^2$ had unit variance and mean equal to its true value of 1.

Although the models are non-nested, the consistency of model estimation arguments for the BIC criteria still apply. TPBF approximates this criteria. Its behaviour in Table 5 across σ_y parameter settings is stable and biased toward the lower-dimensional model. Similarly BF is expected to be biased toward the lower-dimensional model. The setting $\sigma_y = \sqrt{2}$ corresponds to an information ratio, $|I_1|^{-1/2}/|I_2|^{-1/2} = 1$ so that no bias due to the information ratio is expected. We see that the behaviour of BF in this case is comparable to TPBF and that BF is indeed biased toward the lower-dimensional model. When $\sigma_y = 5$, which corresponds to a information ratio that is larger than 1, the bias is even more pronounced, as expected. The bias coming from the information ratio reinforces the bias toward lower-dimensional models.

What is surprising is the case $\sigma_y = 0.1$. Here the information ratio is less than 1. That causes bias toward Model 2 that is strong enough to counter-act the bias toward the simpler model.

TABLE 5: The percentages of times $BF_{12}/(1 + BF_{12})$ (first number) or $BF_{21}/(1 + BF_{21})$ (second number) was larger than 0.5, 0.6, 0.75, and 0.9 when Models 1 and 2 are both correct in Example 5.

	$\sigma_y = 0.1$				$\sigma_y = \sqrt{2}$				$\sigma_y = 5.0$			
	0.50	0.60	0.75	0.90	0.50	0.60	0.75	0.90	0.50	0.60	0.70	0.90
BF	12/88	7/79	3/42	1/13	91/9	86/6	73/3	22/1	96/4	95/2	90/1	69/0
Model prior	99/1	99/1	97/0	92/0	86/14	79/10	60/5	13/2	39/61	15/42	4/21	0/8
FBF $b = 0.01$	76/24	64/15	27/7	6/2	89/11	83/7	66/4	16/1	95/5	93/3	86/2	58/0
FBF $b = 0.5$	63/37	21/16	3/4	0/0	64/36	24/16	4/4	0/0	70/30	33/13	3/4	0/0
PBF $b = 1.5$	52/48	15/21	2/5	0/1	53/47	15/22	3/5	0/1	55/45	16/20	2/5	0/1
PBF $b = 2$	49/51	25/34	9/17	2/5	50/50	26/34	10/17	3/6	52/48	29/32	9/17	2/6
TPBF	90/10	85/6	70/3	19/1	90/10	84/7	69/3	20/1	90/10	85/7	71/4	20/1

4. PARAMETERIZATION AND BIAS FACTORS

As the model prior correction was usually effective at reducing bias and only corrects for the information ratio term in (5), it is clear that for the parameterizations considered, the main reason for bias is the information ratio, not the priors for the parameters. In all of the examples natural parameterizations were chosen. We now show that once alternative parameterizations are allowed, bias for the same model comparison may be viewed as entirely due to the prior ratio, entirely due to the information ratio or a mixture of both.

The determinant of the information for a model after transformation to η is $|I(\eta)| = |\eta'(\theta)|^{-2}|I(\theta)|$ and the prior is $\pi(\eta) = \pi(\theta)|\eta'(\theta)|^{-1}$. Taking products, gives that $|I(\eta)|^{-1/2}\pi(\eta) = |I(\theta)|^{-1/2}\pi(\theta)$. Thus the product of the information and prior ratios is unaffected by reparameterization and the same bias term (5) is expected regardless of how the model is parameterized. Separately, however, information ratios and prior ratios are affected by reparameterization. The parameter transformation $\eta(\theta_1) = \theta_1 \times |I_1(\theta_{10})|^{-1/2}/|I_2(\theta_{20})|^{-1/2}$ gives an information ratio satisfying

$$|I_1(\hat{\eta})|^{-1/2}/|I_2(\hat{\theta})|^{-1/2} \approx |I_1(\eta_{10})|^{-1/2}/|I_2(\theta_{20})|^{-1/2} = 1,$$

where $\eta_{10} = \eta(\theta_{10})$. Alternatively $\eta(\theta_1) = \theta_1 \times \pi(\theta_{10}|1)/\pi(\theta_{20}|2)$ gives a prior ratio which, after transformation, satisfies

$$\pi(\hat{\eta}|1)/\pi(\hat{\theta}|2) \approx \pi(\eta_{10}|1)/\pi(\theta_{20}|2) = 1.$$

Thus, depending upon the parameterization, the bias due to the product of the prior ratio and the information ratio can be viewed as entirely due to the prior ratio, entirely due to the information ratio or a mix of both.

The analysis above implies that the explanations for bias differ even for the same problem. In the case that the prior ratio is 1 and the information ratio is larger than 1 an intuitive explanation for bias is as follows. The effective region of integration when calculating $p(\boldsymbol{x}|m) = \int L_m(\theta)\pi(\theta|m)\, d\theta$ is the region where $L_m(\theta)\pi(\theta|m)$ is large. As $L_m(\theta)$ is usually much more concentrated in its mass than $\pi(\theta|m)$, integration is effectively over the region where $L_m(\theta)$ is large. If this region has relatively large volume, $p(\boldsymbol{x}|m)$ will usually be large but if it has small volume $p(\boldsymbol{x}|m)$ will usually be small, even if $L_m(\theta)$ has a large maximum. Stated another way, if the data support a large subset of parameters poorly in Model 1, its integrated $L_m(\theta)$ can be much larger than a Model 2 for which a small subset of parameters are well supported. As $I_m(\hat{\theta}_m)^{-1}/n$ approximates the covariance matrix of the $\hat{\theta}_m$, an information ratio greater than 1 is indicative of a data set for which parameters are less well supported under Model 1 than 2. Such data sets can be expected to arise in a number of settings including high-variance parameter estimation (Example 1), high overall variability (Example 2), high correlation of parameter estimates (Example 3) and unequal stratification (Example 4).

5. BIAS FACTORS AS A CONSEQUENCE OF PRIOR KNOWLEDGE

In the case that the information ratio is 1, bias due to (5) is a consequence of a prior ratio differing from 1, a situation that we now explore. Consider, for illustration, the example from Section 3.1. Reparameterizing (μ_x, μ_y) as $(\eta_x, \eta_y) = (\sigma_y\mu_x, \mu_y)$ gives an information ratio of 1, so that bias is entirely due to the prior ratio. In Model 1 $\eta_x = 0$ and the prior for η_y is $N(0, 1)$. In Model 2 $\eta_y = 0$ and the prior for $\eta_x \sim N(0, \sigma_y^2)$. If $\mathrm{Var}(\eta_x) = \sigma_y^2 > 1$, then prior knowledge indicates that it is more likely that $\eta_y \approx 0$ than that $\eta_x \approx 0$. Assume that in reality $\eta_x = \eta_y = 0$, so that both models are correct. Then from Table 1 when $\sigma_y = 5$ there is a 94% chance that one will observe $BF > 1$. The direct interpretation that one might ascribe to each of these $BF > 1$ is that it is more

likely that $\eta_x = 0$ and η_y differs from 0 (Model 1) than that η_x differs from 0 and $\eta_y = 0$ (Model 2). This in spite of the fact that prior information suggests it is more likely that $\eta_y \approx 0$ than that $\eta_x \approx 0$.

The seeming paradox arises because what is meant by a "model" in the Bayesian context differs from the frequentist definition above. Model 1 actually is a more specific statement than that $\eta_x = 0$ and η_y differs from 0. Rather the claim is that $\eta_x = 0$ and $\eta_y \sim N(0, 1)$. The reason that there is a 94% chance that one will observe $BF > 1$ is that the Model 1 $N(0, 1)$ prior for η_y places much more mass at the true (but unknown) $\eta_y = 0$ than the Model 2 $N(0, \sigma_y^2)$ prior for η_x. Some would argue that the result makes perfect sense: in the absence of information from the data the prior determines which model is selected.

An alternative perspective is that prior ratios differing from one are a consequence of incompatibility of prior information. Within a model the prior density assigned to θ is uniquely defined. Stated more generally, in the space of probability distributions for the data, the prior density on any particular probability distribution, $p(\boldsymbol{x}|\theta)$, is well defined. The extension of this principle to multiple models is that the density, $\pi(\theta|m)\alpha_m$, associated with any particular probability distribution, $p(\boldsymbol{x}|\theta, m)$, should be uniquely defined. For intersecting models where $p(\boldsymbol{x}|\theta_{10}, 1) = p(\boldsymbol{x}|\theta_{20}, 2)$ the principle of uniquely defined density on probability distributions requires that $\pi(\theta_{10}|1)\alpha_1 = \pi(\theta_{20}|2)\alpha_2$. The use of Bayes factors implicitly assumes $\alpha_1 = \alpha_2$, giving the requirement that $\pi(\theta_{10}|1) = \pi(\theta_{20}|2)$. If prior density on probability distributions for the data is uniquely defined, prior ratios differing from one should not arise.

Following the principle of uniquely defined prior densities thus provides a way of avoiding bias (assuming an information ratio of 1 at points of model intersection). Putting aside the issue of bias the principle is of interest in itself as a principle for specifying prior knowledge in a logically consistent manner. The conventional approach of specifying priors about parameters within models is the source of the problem and leads to the additional difficulty that it is impossible to follow the principle in a transformation invariant way: after transformation $\pi(\eta_{10}|1) = \pi(\theta_{10}|1)|\eta'(\theta_{10})|^{-1}$. One possibility is to require that prior density be uniquely defined for whatever parameterization is under consideration. Practical challenges in implementing the principle for complex models include determining points or even regions of intersection between models.

6. DISCUSSION

The series of examples illustrate how biases with Bayes factors can arise in a number of different settings involving non-nested models. PBF, model priors and, if b is not too small, FBF corrections are effective at reducing bias. The bias is asymptotic: it does not vanish as sample size gets large.

The three adjustments considered were PBF, FBF, and the model prior adjustment. Another class of adjustments to Bayes factors are the intrinsic Bayes factor adjustments of Berger & Pericchi (1996). For illustration consider a subset of these that can be constructed as follows. Bayes factors, $B_{12}(\boldsymbol{x}_l)$ are calculated for subsets, $\boldsymbol{x}_l$, of the data. These subsets of the data are chosen as the smallest subsets satisfying that $0 < p(\boldsymbol{x}_l|m) < \infty$; often they correspond to the individual observations. An intrinsic Bayes factor is then calculated as BF_{12} multiplied by some measure of the middle of the distribution of the $1/B_{12}(\boldsymbol{x}_l)$ over $\boldsymbol{x}_l$, such as the mean or the median. It seems possible that, for some settings, the $B_{12}(\boldsymbol{x}_l)$ will exhibit some of the same biases that would be seen with the full data $\boldsymbol{x}$, in which case multiplying B_{12} by the average $1/B_{12}(\boldsymbol{x}_l)$ might reduce bias. It is unclear, however, that some version of the approach will asymptotically eliminate bias.

In the absence of corresponding likelihood implementations model prior corrections can be difficult to calculate. An appealing feature of PBF or TPBF corrections are that they can easily be approximated without requiring the computational backflips usually required for Bayes factor

calculations (Meng & Wong, 1996; Gelman & Meng, 1998). Any Bayesian software that outputs likelihoods can be used to approximate PBF through the ratio of average exponentiated likelihoods (8) or (9), which are intuitively natural measurements of support for Model 1. By contrast with PBF the natural approximation to FBF is

$$B^{-1}\sum_{k=1}^{B} L_2(\theta_{2k})^{b-1} / B^{-1}\sum_{k=1}^{B} L_1(\theta_{2k})^{b-1} \tag{15}$$

which can be expected to be much less stable. Whereas for $b > 1$ $E([L_m(\theta)]^{b-1}|\boldsymbol{x})$ can be expected to be reasonably well estimated by the average $[L_m(\theta_k)]^{b-1}$, when $b < 1$, the average $L_m(\theta_k)^{b-1}$ will be dominated by the occasional small $L_m(\theta_k)$ associated with a tail posterior value θ_k. Indeed, for b small the approximation (15) becomes similar to the harmonic mean estimator of Newton & Raftery (1994) which has been noted to be an unstable approximation in many cases (Raftery et al., 2007).

Use of TPBF requires a choice of b. With few exceptions the estimates of TPBF were all similar regardless of whether $b = 0.5$, 1.5, or 2.0 was used, as is expected based on (11). As approximation (8) is easily calculated, comparison across multiple choices of b is a useful cross-check as to whether simulation-based approximations (which were not required here) of TPBF work well. Approximations can be expected to do well when $V_k = L_m(\theta_k)^{b-1}/\max_k L(\theta_k)^{b-1}$ is not heavily skewed in either direction. With $b \approx 1$ V_k is skewed toward 1 whereas with b large it is skewed toward 0. Computation with several choices of b to find values giving mean V_k near 0.5 is a reasonable strategy.

When the posterior Bayes factor was first introduced it was criticized on several grounds. One of these was a troubling example in Lindley (1991) but that example was subsequently shown to be a case of Simpson's paradox in Aitkin (1998). More serious is the behaviour for nested models or models of differing dimension. Here uncorrected use of PBF, with $b = 2$ for illustration, is asymptotically equivalent to the following *IC criteria: prefer more complex Model 1 to 2 if $l_1(\hat{\theta}_1) - l_2(\hat{\theta}_2) - P > 0$, but with a small penalty $P = 0.35(d_1 - d_2)$ in place of $(d_1 - d_2)\log(n)/2$ for BIC and $d_1 - d_2$ for AIC. As BIC is, up to higher order terms, equivalent to Bayesian model selection, the issue is related to the criticism that PBF is not Bayesian. The difficulty can be circumvented by using TPBF which, asymptotically, matches results that would be obtained using reasonable normal priors. Viewed from this perspective, the use of PBF provides a simple means, given the availability of a Bayesian software implementation and almost any priors, of obtaining an approximation to a Bayesian approach with those priors being replaced with reasonable normal priors.

APPENDIX

Proof of Theorem 1. As $l_1(\theta_{10}) = l_2(\theta_{20})$, taking logarithms of (4) gives that

$$2\log BF_{12} = 2\log b(\hat{\theta}_1, \hat{\theta}_2) + 2\{l_1(\hat{\theta}_1) - l_1(\theta_{10})\} - 2\{l_2(\hat{\theta}_2) - l_2(\theta_{20})\} + O_p(n^{-1/2}). \tag{A.1}$$

Assuming the regularity conditions for Laplace approximation, standard likelihood theory (cf. Chapter 5 of Lehmann, 1983) gives that

$$2\{l_m(\hat{\theta}_m) - l_m(\theta_{m0})\} = ||V_{mn}||^2 + O_p(n^{-1/2}),$$

where $V_{mn} = n^{-1/2} I_m(\theta_{m0})^{-1/2} l'(\theta_{m0})$. Further $E[l'(\theta_{m0})] = 0$ and $\text{Var}[l'(\theta_{m0})/\sqrt{n}] = I_m(\theta_{m0})$. As

$$l'(\theta_{m0}) = \sum_{i=1}^{n} \frac{\partial}{\partial \theta_j} \log p(x_i|\theta_{m0}, m),$$

it follows that $V_{mn} = n^{-1/2} \sum_{i=1}^{n} g_m(X_i)$, for some $g_m(X_i)$ satisfying that $E[g_m(X_i)] = 0$ and $\text{Var}[g_m(X_i)] = I$. Let $G_{rs} = \text{Cov}[g_1(X_i)_r, g_2(X_i)_s]$. Then the Central Limit Theorem gives that $[V_{1n}^\top, V_{2n}^\top]^\top$ converges in distribution to a $N(0, \Sigma)$ distribution where

$$\Sigma = \begin{bmatrix} I & G \\ G^\top & I \end{bmatrix}.$$

Substituting in (A.1) and using that $b(\hat{\theta}_1, \hat{\theta}_2) = b(\theta_{10}, \theta_{20}) + O_p(n^{-1/2})$ gives that

$$P(2 \log BF_{12} > w) \to P(||V_1||^2 - ||V_2||^2 + 2 \log b(\theta_{10}, \theta_{20}) > w),$$

where $[V_1^\top, V_2^\top]^\top \sim N(0, \Sigma)$. Suppose that the singular value decomposition for G is $G = UDW^\top$, where U and V are orthogonal and D is diagonal. For $Z_1 = U^\top V_1$ and $Z_2 = W^\top V_2$, $||V_1||^2 - ||V_2||^2 = ||Z_1||^2 - ||Z_2||^2$ and $[Z_1^\top, Z_2^\top]^\top \sim N(0, \Sigma_z)$ where

$$\Sigma_z = \begin{bmatrix} I & D \\ D & I \end{bmatrix}.$$

The labelling of Z_1 and Z_2 is arbitrary so that the distribution of $||Z_1||^2 - ||Z_2||^2$ must be the same as $||Z_2||^2 - ||Z_1||^2$. In summary

$$P(2 \log BF_{12} > w) \to P(W_0 + 2 \log b(\theta_{10}, \theta_{20}) > w)$$

where W_0 has a symmetric distribution around 0. As $BF_{12} = 1/BF_{21}$,

$$\begin{aligned} P(BF_{21} > t) &= P(BF_{12} < 1/t) \\ &= P\{2 \log BF_{12} < -2 \log(t)\} \\ &= P\{W_0 < -2 \log(t) - 2 \log b(\theta_{10}, \theta_{20})\} + o(1) \\ &= P\{W_0 > 2 \log(t) + 2 \log b(\theta_{10}, \theta_{20})\} + o(1) \\ &= P\{W_0 + 2 \log b(\theta_{10}, \theta_{20}) > 2 \log(t) + 4 \log b(\theta_{10}, \theta_{20})\} + o(1) \\ &= P\{2 \log BF_{12} > 2 \log(t) + 4 \log b(\theta_{10}, \theta_{20})\} + o(1) \\ &= P\{BF_{12} > t b(\theta_{10}, \theta_{20})^2\} + o(1). \end{aligned}$$

■

ACKNOWLEDGEMENTS

This research was supported by a Discovery Grant from the Natural Sciences and Engineering Research Council of Canada. I thank two anonymous reviewers for valuable comments which greatly aided the manuscript, in particular, leading to discussion of how reasons for bias will be viewed differently depending on the parameterization considered.

BIBLIOGRAPHY

Aitkin, M. (1998). Simpson's paradox and the Bayes factor. *Journal of the Royal Statistical Society, Series B*, 60, 269–270.

Aitkin, M. (1991). Posterior Bayes factors. *Journal of the Royal Statistical Society, Series B*, 53, 111–142.

Berger, J. O., Bernardo, J. M., & Sun, D. (2009). The formal definition of reference priors. *The Annals of Statistics*, 37, 905–938.

Berger, J. O. & Pericchi, L. J. (2001). Objective Bayesian methods for model selection: Introduction and comparison. *Institute of Mathematical Statistics Lecture Notes-Monograph Series*, 38, 135–207.

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

Burnham, K. P. & Anderson, D. R. (2013). *Model Selection and Inference: A Practical Information-Theoretic Approach*. Springer, New York.

Gelman, A. & Meng, X. (1998). Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical Science*, 13, 163–185.

Ghosh, M. (2011). Objective priors: An introduction for frequentists. *Statistical Science*, 26, 187–202.

Gilks, W. R., Richardson, S., & Speigelhalter, D., editiors. (1996). *Practical Markov Chain Monte Carlo*. Chapman and Hall, New York.

Jeffreys, H. (1961). *Theory of Probability*, 3rd ed., Oxford University Press, London.

Kadane, J. B. & Lazar, N. A. (2004). Methods and criteria for model selection. *Journal of the American Statistical Association*, 99, 279–290.

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

Kass, R. E., Tierney, L., & Kadane, J. B. (1990). The validity of posterior expansions based on Laplace's method. In *Bayesian and Likelihood Methods in Statistics and Econometrics*, Hodges, J. S., Press, S. J., & Zellner A., editiors. Elsevier, Amsterdam, pp. 473–488.

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.

Kass, R. E. & Wasserman, L. (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, 91, 1343–1370.

Kass, R. E. & Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association*, 90, 928–934.

Lehmann, E. L. (1983). *Theory of Point Estimation*. John Wiley and Sons, New York.

Lehmann, E. L. (1951). A general concept of unbiasdness. *The Annals of Mathematical Statistics*, 22, 587–592.

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

Lindley, D. V. (1991). Discussion of "Posterior Bayes factors", by M. Aitkin. *Journal of the Royal Statistical Society, Series B*, 53, 130–131.

Meng, X. & Wong, W. H. (1996). Simulating ratios of normalizing constants via a simple identity: A theoretical exploration. *Statist. Sinica*, 6, 831–860.

Newton, M. A. & Raftery, A. E. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society, Series B*, 56, 3–48.

Neyman, J. & Pearson, E. S. (1936). Contributions to the theory of testing statistical hypotheses. I. Unbiased critical regions of type A and type A_1. *Statistical Research Memoirs*, 1, 1–37.

O'Hagan, A. (1995). Fractional Bayes factors for model comparison. *Journal of the Royal Statistical Society, Series B*, 57, 99–138.

Raftery, A. E., Newton, M. A., Satagopan, J. M., & Krivitsy, P. N. (2007). Estimating the integrated likelihood via posterior simulation using the harmonic mean identity. *Bayesian Statistics*, 8, 1–45.

Schwarz, G. E. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6, 461–464.

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

Received 24 February 2017
Accepted 13 April 2017