

The Effect of Priors on Approximate Bayes Factors from MCMC Output^a

Kevin A. Clarke
University of Michigan, Department of Political Science
505 South State Street, Ann Arbor, MI 48109-1045
kclarke@umich.edu
<http://www.umich.edu/~kclarke>

October 4, 2000

Abstract

The MCMC approach to calculating approximate Bayes factors is considered. The calculation, consisting of a log-likelihood, a prior, and a posterior, presents an excellent opportunity to observe directly the effects of priors on Bayes factors. Three empirical examples demonstrate that Bayes factors are sensitive to a *combination* of the prior variance and the difference in the number of parameters between the rival models.

^aI thank Susan Murphy for helpful discussions and Paul Huth, Christopher Gelpi, D. Scott Bennett, Dan Reiter, and Al Stam for sharing their data.

1 Introduction

Bayesian inference in political science has undergone a revival of late. Advances in simulation techniques that make the math of Bayesian analysis tractable, along with decreases in computing costs, have combined to make Bayesianism attractive (see (Western and Jackman 1994) and (Jackman 2000)). The resurgence of Bayesian inference in the social sciences has generated renewed interest in the calculation and use of Bayes factors (the ratio of marginal likelihoods for two models—see Section 4). This revival began, in the social sciences, with Adrian Raftery’s (1995) work on model selection and uncertainty in sociology. The use of Bayes factors in political science followed in articles by Bartels (1997) and Smith (1999).

Interest in Bayes factors has centered on two properties social scientists consider attractive. The first is that Bayes factors overcome the difficulties that p-values present (Raftery 1995). The second is that Bayes factors overcome the difficulty presented by the posterior odds ratio, namely that the researchers must specify prior distributions (Lavine and Schervish 1999), and therefore Bayes factors can be used to test hypotheses “objectively.” I demonstrate through the use of the MCMC approximation to Bayes factors and a few simple, empirical examples that these claims are not always borne out.¹ The results show that Bayes factors can be highly sensitive to a combination of the prior variance and the difference in the number of parameters between the rival models. Bayes factors, then, do not completely escape the problem of priors, and the uncritical use of Bayes factors can be misleading. The remainder of the paper addresses some general issues regarding the interpretation and use of Bayes factors.

In Section 2, I briefly review some of the more telling criticisms of the use of classical p-values. In Section 3, I review the argument that Bayes factors are as necessary to Bayesian inference as p-values are to classical inference. In Section 4, I discuss calculating Bayes factors from Markov Chain Monte Carlo (MCMC) output. In Section 5, I present three illustrative examples, two of which are drawn from the literature of international relations, and in Section 6, I discuss the interpretation of Bayes factors.

¹I write of *the* MCMC approximation for Bayes factors for the sake of convenience. There are, in fact, many MCMC approximations (see Han and Carlin 2000).

2 The Problem with P-Values

The problems with p-values have been the subject of a vast amount of research over the years. Useful discussions of these problems can be found in Edwards, Lindman, and Savage (1963), Berger and Sellke (1987), Raftery (1995), and Gill (1999). Of all the criticisms of p-values, the most damaging is the claim that p-values do not provide a measure of the amount of evidence in favor of the null hypothesis. Berger and Sellke (1987) provide a simple example where a Bayesian gives a prior probability of 0.5 to both H_0 and H_1 . They demonstrate that if $n = 1000$ and $t = 2.576$, one can classically reject H_0 at $p = .01$ although the actual probability of the null given the data is 0.53. The evidence, then, favors the null. Most political scientists, working from the p-value, would make an incorrect decision regarding the null in this case.

An interesting, alternative approach to making a similar point comes from Schervish (1996). A measure of support should demonstrate the logical property of *coherence*,

If hypothesis H implies hypothesis H' , then there should be at least as much support for H' as there is for H .

For example, to be coherent, the measure of support for $H' : M \leq 2$ must be at least as large as the measure of support for $H : M \leq 3$. Schervish demonstrates that, in general, p-values do not have this property and therefore fail as measures of support.

As a numerical example, Schervish (1996) defines two interval hypotheses: $H : M \in [-.5, .5]$ and $H' : M \in [-.82, .52]$. Note that H implies H' .² The p-value for an interval hypothesis is given by Schervish (see Lehmann (1986) for more on tests of interval hypotheses),

$$p_{\mu_1, \mu_2}(x) = \begin{cases} \Phi(x - \mu_1) + \Phi(x - \mu_2) & \text{if } x < .5[\mu_1 + \mu_2] \\ \Phi(\mu_1 - x) + \Phi(\mu_2 - x) & \text{if } x \geq .5[\mu_1 + \mu_2]. \end{cases} \quad (1)$$

If $x = 2.18$, the p-values calculated from (1) are,

$$\begin{aligned} p_{-.50, .5}(2.18) &= \Phi(-.50 - 2.18) + \Phi(.5 - 2.18) = .0502 \\ p_{-.82, .52}(2.18) &= \Phi(-.82 - 2.18) + \Phi(.52 - 2.18) = .0498. \end{aligned} \quad (2)$$

²If the parameter is between $[-.5, .5]$, then it must be between $[-.82, .52]$.

The p-value is incoherent because there is more support for H than for H' , even though H implies H' .

3 The Necessity of Being Comparative

The arguments presented in Section 2 clearly demonstrate the need for Bayesian inference. Why, though, do we need Bayes factors? After all, the putative strength of the Bayesian account of inference lies in its ability to provide a direct measure of support for a hypothesis. Is it not good enough to state that a hypothesis has a certain level of support?

The answer to the question above would certainly be yes...if it were possible to make such a statement. Unfortunately, Bayesian inference suffers from a logical flaw that prevents making any such statements for a single hypothesis.

Briefly, Bayes' Theorem states:

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)}. \quad (3)$$

That is, the probability of the hypothesis given the data is proportional to the prior probability of the hypothesis multiplied by the likelihood of the data given the hypothesis.³ The problem lies with the normalizing constant of the posterior density, $P(D)$.

The denominator of equation (3) is the theorem of total probability, which states in general,

$$\begin{aligned} &\text{If } P(a_i \vee \dots \vee a_n) = 1, \text{ and } a_i \vdash \neg a_j \text{ for } i \neq j, \text{ then} \\ &\text{for any } b, P(b) = P(b|a_i)P(a_i) + P(b|\neg a_i)P(\neg a_i).^4 \end{aligned} \quad (4)$$

When we apply (4) to equation (3), the denominator becomes,

$$P(D) = P(D|H)P(H) + P(D|\neg H)P(\neg H). \quad (5)$$

³For what follows, the term "hypothesis" could easily be replaced with the term "model."

⁴ \vdash means "implies," \vee is a disjunction, and \neg means "not"

The troublesome aspect of (5) is the $\neg H$ in the second term. If only two possible hypotheses existed, we could evaluate equation (5) without any problem. In an actual research situation, however, there are bound to be more than two possible hypotheses. In fact, an infinite number of hypotheses are possible making the term $\neg H$ a disjunction of alternative hypotheses, one of which must be logically true. The question, then, is how do we evaluate $P(D|\neg H)$?

In the past, this dilemma was finessed philosophically by Shimony's (1970) "catchall hypothesis," where $\neg H$ is a disjunction of seriously considered alternative theories and the "catchall" hypothesis,

$$\neg H \equiv H_1 \vee H_2 \vee H_3 \vee H_4 \vee H_{n-1} \vee H_c. \quad (6)$$

H_c contains all possible hypotheses not explicitly considered by the researcher.

The problem as Earman (1992) points out, is that "the catchall H_c says, in effect, that some as yet uninvented theory is true." Wesley Salmon (1990) agrees, arguing, "Among the hypotheses hidden in the catchall are some that, in conjunction with present available background knowledge, entail the present evidence." Calculating $P(D|\neg H)$, therefore, is not possible. Without full knowledge of all the possible competing theories, the likelihood $P(D|\neg H)$ is intractable.

If $P(D|\neg H)$ cannot be evaluated, then Bayesian induction is not possible.⁵ If we cannot know the likelihoods of all the possible theories, particularly those in the "catchall" hypothesis, then we cannot calculate the denominator of Bayes' Theorem. Salmon (1990) proposes a solution to this impasse that alters the goal of Bayesian analysis from measuring direct support to measuring relative support. Salmon demonstrates that if we restrict the theories under consideration to two, we can avoid the troublesome $P(D|\neg H)$ term. Spelling out Bayes' Theorem for each of the two competing theories, we can see the denominators are equal:

$$P(H_1|D) = \frac{P(D|H_1)P(H_1)}{P(D|H_1)P(H_1) + P(D|H_2)P(H_2)} \quad (7)$$

$$P(H_2|D) = \frac{P(D|H_2)P(H_2)}{P(D|H_1)P(H_1) + P(D|H_2)P(H_2)}. \quad (8)$$

⁵At least it is not possible without assuming a radical stance toward the objectivity of science.

We can then take the ratio,

$$\frac{P(H_1|D)}{P(H_2|D)} = \frac{P(D|H_1)P(H_1)}{P(D|H_2)P(H_2)}. \quad (9)$$

The resulting ratio is known as the posterior odds ratio. It does not tell us the degree to which the data support hypothesis 1 or hypothesis 2. Rather, the ratio tells us the degree to which the data support hypothesis 1 over hypothesis 2.

The Bayes factor is an attempt to create a comparative measure like the posterior odds ratio, without having to specify priors (Lavine and Schervish 1999).

4 Bayes Factors

Rewriting equation (9) slightly,

$$\frac{P(H_1|D)}{P(H_2|D)} = \left[\frac{P(D|H_1)}{P(D|H_2)} \right] \left[\frac{P(H_1)}{P(H_2)} \right], \quad (10)$$

the first factor on the right-hand side of (10) is defined as the Bayes factor (the ratio of marginal likelihoods), while the second factor is the prior odds ratio. Equation (10), then, corresponds to,

$$\text{Posterior odds} = \text{Bayes factor} \times \text{Prior odds}. \quad (11)$$

I defined the Bayes factor in (10) as the ratio of marginal likelihoods. The marginal likelihood for H_1 is,

$$P(D|H_1) = \int P(D|\theta, H_1)P(\theta|H_1)d\theta, \quad (12)$$

where θ is the parameter vector, $P(D|\theta, H_1)$ is the likelihood, and $P(\theta|H_1)$ is the prior.

If we assign equal priors to the models in (10), the prior odds ratio equals 1 and drops out. The posterior odds ratio then equals the Bayes factor.

When the Bayes factor is greater than 1, the data favor H_1 over H_2 . When the Bayes factor is less than 1, the reverse is true. Raftery (1994), following Jeffreys (1961), proposes the following “rules of thumb” for interpreting twice the logarithm of the Bayes factor,

$0 \leq 2 \log(BF) \leq 2.2$	Very weak evidence for H_1
$2.2 \leq 2 \log(BF) \leq 5$	Weak to moderate evidence for H_1
$5 \leq 2 \log(BF) \leq 10$	Moderate to strong evidence for H_1
$2 \log(BF) > 10$	Decisive evidence for H_1

One of the attractive benefits of Bayes factors, noted by Raftery (1995), is that posterior probabilities for each hypothesis can be calculated from them. If H_i are the hypotheses being compared with the null (H_0), B_{i0} are the corresponding Bayes factors, and α_i are the prior odds, then,

$$P(H_i|D) = \frac{\alpha_k \beta_{k0}}{\sum_{i=0}^K \alpha_i \beta_{i0}} \quad (13)$$

are the posterior probabilities of the models given the data.

While the theory and use of Bayes factors appear straightforward⁶, calculating the integrated or marginal likelihoods necessary for Bayes factors is far from straightforward.

4.1 Bayes factors from MCMC Output

The MCMC approach to calculating Bayes factors was developed by Chib (1995) and has been used in political science by Smith (1999).⁷ The method is of interest because of its applicability in a wide range of settings (Han and Carlin 2000). The method is also of interest because, as we shall see, the transparency of the method makes it particularly useful for assessing the effects of priors on Bayes factors.

If we rewrite equation (3) in terms of the data and a coefficient vector θ , Bayes’ Theorem is,

⁶See the discussion in Section 6.

⁷This discussion closely follows Chib (1995), although there are minor differences in notation.

$$p(\theta|y) = \frac{f(y|\theta)p(\theta)}{m(y)}, \quad (14)$$

where $p(\theta|y)$ is the posterior density, $f(y|\theta)$ is the likelihood function for the given model, $p(\theta)$ is the prior density, and $m(y)$ is the normalizing constant of the posterior density.

Solving equation (14) for the normalizing constant gives what Chib refers to as the “basic marginal likelihood identity”,

$$m(y) = \frac{f(y|\theta)p(\theta)}{p(\theta|y)}. \quad (15)$$

The natural logarithm of (15) presents a computationally convenient expression,

$$\ln m(y) = \ln f(y|\theta) + \ln p(\theta) - \ln p(\theta|y). \quad (16)$$

Estimating (16) requires choosing a θ^* from the posterior density of θ and evaluating all three terms at θ^* . Chib argues that the choice of θ^* is not critical as long as it is taken from a point of high density.⁸ The estimate of the marginal likelihood is therefore,

$$\ln \hat{m}(y) = \ln f(y|\theta^*) + \ln p(\theta^*) - \ln \hat{p}(\theta^*|y). \quad (17)$$

If we calculate $\ln \hat{m}(y)$ for two rival models, M_f and M_g , the estimate of the Bayes factor is given by,

$$\hat{B}_{fg} = \exp\{\ln \hat{m}(y|M_f) - \ln \hat{m}(y|M_g)\}. \quad (18)$$

Estimating the Bayes factor, then, requires simply evaluating the log-likelihood, the prior, and the posterior density at θ^* for two models. While

⁸The estimations in Section 5 use the posterior mean, which in those cases, is a point of high density. Calculating θ^* is discussed in the following paragraphs.

the first two terms in each marginal likelihood are quite easy to calculate, estimating the posterior densities is more difficult and requires some knowledge of MCMC methods.

I now digress briefly in order to review the MCMC approach to Bayesian estimation. What follows is simply meant as a reminder and not as a full introduction to MCMC methods. A very readable introduction to MCMC methods can be found in Jackman (2000).

For the sake of simplicity, I only consider MCMC estimation for probit models. Albert and Chib (1993) discuss MCMC estimation of both binary and polychotomous response data for a variety of link functions.

The MCMC approach begins by defining a probit model through a latent variable, z_i ,

$$z_i \sim N(x_i'\theta, 1); \quad y_i = I(z_i > 0). \quad (19)$$

Notice that if z_i were known, equation (19) would be a linear regression model,

$$Z = X\theta + \epsilon. \quad (20)$$

By standard regression results,

$$\begin{aligned} \theta|Z \text{ is distributed } & N(\hat{\theta}, (X'X)^{-1}) \\ \text{where } \hat{\theta} &= (X'X)^{-1}(X'Z). \end{aligned} \quad (21)$$

Of course, z_i is a latent variable and it is unknown. We can use Gibbs sampling (see Casella and George 1992), however, to impute our unknowns, z_i and θ . Given proper prior information in the form of a multivariate normal distribution,

$$\theta \sim N(a, A), \quad (22)$$

Albert and Chib (1993) give the *estimated* posterior distribution of θ :

$$\theta|y, Z \sim N(\hat{\theta}_z, B) \tag{23}$$

where $\hat{\theta}_z = (A^{-1} + X'X)^{-1}(A^{-1}a + X'Z)$
and $B = (A^{-1} + X'X)^{-1}$.

Jackman (2000) points out that with an uninformative prior, $\hat{\theta}_z = (X'X)^{-1}(X'Z)$ and $B = (X'X)^{-1}$, which is equal to running the regression of z_i on X .

The complete conditional densities needed for running the Gibbs sampler, then, are:

$$p(\theta|y, z) = \phi(\theta|\hat{\theta}_z, B) \text{ and} \tag{24}$$

$$p(z_i|y, \theta) \propto \phi(z_i|x_i'\theta, 1)I[0, \infty] \text{ if } y_i = 1 \tag{25}$$

$$p(z_i|y, \theta) \propto \phi(z_i|x_i'\theta, 1)I[-\infty, 0] \text{ if } y_i = 0,$$

where ϕ is the normal density.

The output from the Gibbs sampler are draws from the distribution of θ , $\theta^{(g)}$, and draws from the distribution of z , $z^{(g)}$. The $\theta^{(g)}$ are averaged to compute θ^* . That is,

$$\theta^* = \frac{\sum_{g=1}^G \theta^{(g)}}{G}. \tag{26}$$

The $z^{(g)}$ are plugged into equation (23) to compute $\hat{\theta}_z^{(g)}$, the estimated posterior mean of θ .

Following our digression, we are now in a position to estimate the posterior density in equation (17), $\ln \hat{p}(\theta^*|y)$. In simple terms, the conditional probabilities used in the Gibbs sampler are $p(\theta|y, z)$ and $p(z|y, \theta)$. While we have $p(\theta|y, z)$, calculating the marginal likelihood calls for $p(\theta|y)$. Therefore, we need to integrate z out of $p(\theta|y, z)$.

If we can write the posterior distribution as,

$$p(\theta|y) = \int p(\theta|y, z)p(z|y)dz, \tag{27}$$

then, the estimate of $p(\theta|y)$ at θ^* , using the technique of monte carlo integration (see Drakos 1995), is,

$$\hat{p}(\theta^*|y) = G^{-1} \sum_{g=1}^G p(\theta^*|y, z^{(g)}). \quad (28)$$

Equation (28) works because $z^{(g)}$ is a draw from $p(z|y)$. Plugging the results from (23) into (28), we get,

$$\hat{p}(\theta^*|y) = G^{-1} \sum_{g=1}^G \phi(\theta^*|\hat{\theta}_z^{(g)}, B). \quad (29)$$

The marginal likelihood for a probit model, combining equations (17, 22, and 29), is therefore,

$$\ln \hat{m}(y) = \ln f(y|\theta^*) + \ln \phi(\theta^*|a, A) - \ln \left\{ G^{-1} \sum_{g=1}^G \phi(\theta^*|\hat{\theta}_z^{(g)}, B) \right\}. \quad (30)$$

5 Empirical Examples

Equation (30) may be estimated by any number of programs. Smith (1999) did his estimation with Gauss. I used WinBUGs to produce the estimates of θ and z and then read the results into STATA to calculate $\ln \hat{m}(y)$.

5.1 Example 1

My first example employs data that Chib (1995) used for illustrative purposes. Models f and g each consist of a single binary covariate, and they share a binary dependent variable y . Following Chib, I specified a multivariate normal prior with the mean of each parameter equal to 0.75. Instead of specifying a single covariance matrix as Chib did, I calculated the Bayes factors for prior variances of 100, 1000, and 10000. The purpose of this small simulation is to assess the impact of the prior variance on the Bayes factor.

The marginal likelihoods for each model and associated Bayes factors are in Table 1.⁹

[Table 1 about here.]

As is quite clear from the table, changing the prior variance had no effect on the Bayes factor. For each variance, the data provide moderately strong evidence for model f . Notice, however, that increasing the prior variance decreases the marginal likelihoods for both models. Of the three terms in equation (30), the change in the variance affects neither the log-likelihood, $\ln f(y|\theta^*)$, nor the log of the estimated posterior, $\ln \left\{ G^{-1} \sum_{g=1}^G \phi(\theta^*|\hat{\theta}_z^{(g)}, B) \right\}$. The change in variance does, however, affect the log of the prior, $\ln \phi(\theta^*|a, A)$. As the prior variance (A) increases, the log of the prior decreases. This change affects the prior of both marginal likelihoods *equally*. The value of the Bayes factor therefore remains unchanged. The same cannot be said of the next example.

5.2 Example 2

My second example employs data from Huth, Gelpi, and Bennett (1993). In that article, Huth and his co-authors attempt to test a model of structural realism against a model of rational deterrence theory.¹⁰ The authors conceptualize structural realism in terms of the amount of uncertainty created by the structure of the international system. To connect the amount of uncertainty in the system to actual decisions taken by state leaders, the authors interact uncertainty with the risk propensities of these decision-makers. When uncertainty is high, risk-acceptant leaders will pursue policies that might spark armed conflict, while risk-averse leaders will likely be more cautious (Huth, Gelpi, and Bennett 1993). Structural realism, then, is operationalized by two composite measures of uncertainty (size and capability diffusion), a measure of risk propensity, and two interaction terms (one for each measure of uncertainty).

⁹To aid later comparisons, the Bayes factors are reported on a scale of twice the natural logarithm of the Bayes factor.

¹⁰The results of the Huth analysis are not germane to this particular discussion and hence go unreported. Interested readers are referred to Huth, Gelpi, and Bennett (1993) and the analyses in Clarke (2000).

As for rational deterrence theory, Huth, Gelpi, and Bennett (1993) argue that, “the credibility of the threat is the primary determinant of deterrence success or failure.” Credibility is affected by the balance of military capabilities, the interests at stake for the states involved, the past dispute behavior of the states, and whether either state is engaged in another dispute at the same time. Deterrence is more likely to fail as the balance of capabilities and the interests at stake shift toward the challenger. Deterrence is also more likely to fail if the defender has backed down in a previous dispute or is engaged in a dispute elsewhere.

The structural realist model contains five covariates and the rational deterrence model contains eight. As in the previous example, I use a multivariate normal prior with the mean of each parameter set to zero and a range of prior variances. The results are in Table 2.

[Table 2 about here.]

Unlike example 1, the prior variances here have a major effect on the Bayes factors. By increasing the prior variance, the Bayes factor moves from moderately strong evidence in favor of the rational deterrence model to moderately strong evidence in favor of the structural realist model. Again, the culprit is not the log-likelihood or the log of the posterior, but rather the log of the prior. The values are in Table 3.

[Table 3 about here.]

As the prior variance increases, the log of the prior decreases, thereby decreasing the marginal likelihood for both models.¹¹ The important difference between examples 1 and 2 is the difference in the number of estimated parameters between the rival models. In this example, the increasing variance has a greater effect on the prior of the model with the greater number of covariates, the rational deterrence model. Increasing the variance therefore favors the “simpler” structural realist model.

If we recalculate the values in Table 3 but this time dropping the last three covariates from the rational deterrence model (making the number of estimated parameters equal), we find that the increasing variance affects the two priors equally. In fact, the priors are just about equal despite the fact that θ^* differs for the two models. The results are in Table 4.

¹¹To be perfectly clear, as A increases, $\ln \phi(\theta^*|a, A)$ decreases.

[Table 4 about here.]

The danger, then, is in the interaction between the prior variance and the difference in the degrees of freedom between the two models. The same finding is borne out in example 3.

5.3 Example 3

My third example employs data from Reiter and Stam (1998). Reiter and Stam are interested in tracing the effect of political structure on war outcomes; that is, they are interested in determining why democracies win more wars than nondemocracies. Is it because democracies are intrinsically more effective at waging war, or because democracies are more careful about the decision to initiate conflict? The argument in favor of the former is that it is easier for democracies to rally their societies behind a war effort and that democratic armies fight “with greater initiative and better leadership than do the armies of other kinds of states” (Reiter and Stam 1998). The argument in favor of the latter is that democratic leaders face greater post-defeat political consequences than do other kinds of states and therefore initiate wars only when the likelihood of victory is high.

In answering this question, Reiter and Stam estimate five models, one of which corresponds to a realist model of war outcomes and one which corresponds to a model that reflects the effects of regime type and the decision to initiate war. The realist argument is that war outcomes are best explained by the distribution of capabilities across combatants, the quality of the militaries involved, strategy choice, terrain, and the effects of allies. The non-realist argument focuses on the regime type of the combatants and the decision to initiate war.

The non-realist model contains three covariates and the realist model contains nine. I again use a multivariate normal prior with the mean of each parameter set to zero and a range of prior variances. We should expect that as the prior variance increases, the Bayes factors should increasingly favor the model with the fewest covariates, i.e. the non-realist model. The results are in Table 5.

[Table 5 about here.]

The results confirm our expectation. The Bayes factors move in favor of the “simpler” non-realist model as the prior variances increase. Unlike example 2, however, the change in the Bayes factor does not affect our inference

regarding which model the data support. The realist model is by far the superior model, but it is unlikely that we would have needed Bayes factors to tell us this.¹²

5.4 Discussion

The idea that Bayes factors are sensitive to the interaction between the prior variance and dimensions of the competing models is not new. Kass and Raftery (1995) mention “Bartlett’s paradox,” which concerns the deleterious effects of using large variances in an attempt to make a prior “noninformative” (Bartlett 1957). The exact connection between the prior variance and the dimensions of the models, however, remains unclear in their discussion.

In an effort to control the size of the prior variance, Raftery (1994) proposes a Laplace approximation for Bayes factors, which included a reference set of proper priors for generalized linear models. These reference priors, denoted by φ , correspond closely to the priors used in the above examples. The results from Raftery’s GLIB (Generalized Linear Bayesian Modeling) are in Table 6.¹³

[Table 6 about here.]

The approximation used by the GLIB software, like the MCMC approximation, fails to control for the difference in the number of estimated parameters between the models and hence the reference priors have large effects.

Given the above, our examples are both suggestive and disturbing. In example 1, the Bayes factors are unaffected by the prior variance, but only because the number of covariates in the competing models are equal. This condition is unlikely to hold in practice very often *and will never hold when the competing models are nested*. The Bayes factors are affected by the prior variance in example 3, but the inference we would make is not affected because the non-realist model garners almost no support from the data. In the most interesting example, example 2, the rival models do not have the same number of covariates and one model is not obviously superior to the other. In this case, not only are the Bayes factors affected by the prior variance, but the inference we would draw from them is affected as well. We can trust

¹²See analyses in Clarke (2000).

¹³The Bayes factors for the Chib data are unaffected by φ and therefore go unreported.

Bayes factors in cases like example 2 only if the dimensions of the models are equal or if extensive sensitivity analysis has been performed and reported.

Granted that the examples provided here are extreme, the results remain unsettling. Claims made on behalf of the “objectivity” of Bayes factors (see Lavine and Schervish 1999) are not borne out. Care must be taken in the use of Bayes factors and sensitivity analyses should be reported.

It may be argued that the objectivity of Bayes factors may be salvaged by removing the specification of the prior from the researcher (Kass and Raftery 1995). One such suggestion is the Bayesian Information Criteria (BIC) approximation used by Raftery (1995). For a null model with no independent variables, M_0 , against a model of interest, M_k , the BIC approximation is,

$$BIC_k = -\chi_{k0}^2 + p_k \log n, \quad (31)$$

where χ_{k0}^2 is the likelihood ratio test statistic for testing M_0 against M_k , and p_k is the degrees of freedom associated with the test. For two models of substantive interest, M_k and M_j , twice the log of the Bayes factor is approximately equal to the difference in their BIC approximations:

$$2 \log B_{jk} \approx BIC_k - BIC_j. \quad (32)$$

The BIC approximation, however, is not really free of priors. Rather, the prior is set and hence objective only in the sense that the researcher cannot affect it. The prior for the BIC approximation is multivariate normal with mean $\hat{\theta}_j$ and covariance matrix i_j^{-1} , where $\hat{\theta}_j$ is the vector of maximum likelihood estimates and i_j is the expected Fisher information for one observation (Bartels 1997). The prior distribution therefore “contains the same amount of information as would, on average, a single observation” (Raftery 1995). The major benefit of the BIC approximation is that it includes the BIC penalty for the number of parameters being estimated. Assessing how much this penalty controls the effects of the prior variance is not immediately apparent as the prior cannot be easily changed and hence is not immediately conducive to simulation.

Bayes factors for our three examples calculated using the BIC approximation are in Table 7.

[Table 7 about here.]

The Bayes factor for the Chib data in Table 7 is extremely close to the Bayes factors for the Chib data reported in Table 1. This isomorphism is only to be expected, as we know that the prior variance will have little effect on Bayes factor when the rival models have the same number of parameters. The BIC approximated Bayes factor for the Reiter and Stam data also corresponds to those reported in Table 5. This result was also to be expected given the disparity between the models. The BIC approximated Bayes factor for the Huth *et al.* data corresponds to the *low* end of the prior variances reported in Tables 2 and 6. Unfortunately, we cannot conclude that this value is correct until we have a better feel for the effects of the BIC penalty.

Unless special circumstances, such as those found in examples 1 and 3, pertain, Bayes factors are sensitive to the interaction between the prior variance and the difference in the number of parameters. Sensitivity analysis is therefore a must.

6 Interpreting Bayes Factors

Up to this point, I have used the standard interpretation of Bayes factors given by Kass and Raftery (1995), “The Bayes factor is a summary of the evidence provided by the data in favor of one scientific theory[...]as opposed to another.” Often this is taken to mean that Bayes factors give us a measure of support for one model over another. A simple example from Lavine and Schervish (1999) contradicts this interpretation.¹⁴

A coin is known to either two-headed, fair, or two-tailed. There are six hypothesis about θ , the probability of heads:

$$\begin{aligned} H_1 : \theta = 1 & \quad H_2 : \theta = \frac{1}{2} & \quad H_3 : \theta = 0 \\ H_4 : \theta \neq 1 & \quad H_5 : \theta \neq \frac{1}{2} & \quad H_6 : \theta \neq 0 \end{aligned}$$

The priors on these hypotheses are,

$$\begin{aligned} P(H_1) = 0.01 & \quad P(H_2) = 0.98 & \quad P(H_3) = 0.01 \\ P(H_4) = 0.99 & \quad P(H_5) = 0.02 & \quad P(H_6) = 0.99^{15} \end{aligned}$$

¹⁴While the example is substantially the same, my notation is quite different.

Suppose the coin is tossed four times and four heads resulted. The posterior probability of H_1 is,

$$P(H_1|X = 4) = \frac{P(X = 4|H_1)P(H_1)}{P(X = 4|H_1)P(H_1) + P(X = 4|H_2)P(H_2)} = 0.14.$$

The posterior probability of H_4 is,

$$P(H_4|X = 4) = 1 - P(H_1|X = 4) = 0.86.$$

The posterior odds ratio in favor of H_4 is then,

$$\frac{P(H_4|X = 4)}{P(H_1|X = 4)} = 6.14. \quad (33)$$

From the posterior odds, we can solve for the Bayes factor. Note that the Bayes factor in equation (10) may be rewritten as the posterior odds over the prior odds,

$$\left[\frac{P(H_4|D)}{P(H_1|D)} \right] = \frac{\left[\frac{P(H_4|D)}{P(H_1|D)} \right]}{\left[\frac{P(H_4)}{P(H_1)} \right]} \quad (34)$$

The Bayes factor in favor of H_4 is then,

$$\left[\frac{P(H_4|D)}{P(H_1|D)} \right] = \frac{6.14}{\frac{0.99}{0.01}} = 0.062. \quad (35)$$

The standard interpretation of the result in equation (35) is that the data offer very little support for H_4 . The posterior odds ratio in equation (33), however, tells us that there is great deal of support for H_4 . The Bayes factor, then, cannot be interpreted as a direct measure of support.

What then does the Bayes factor measure? Lavine and Schervish (1999) state, “What the Bayes factor actually measures is the *change* in the odds in favor of the hypothesis when going from prior to the posterior.” In the example just given, the small Bayes factor indicates that the data substantially lowers the probability of H_4 from 0.99, the prior, to 0.86, the posterior. The Bayes factor cannot be interpreted as indicating that H_4 is unlikely.

¹⁵ $P(H_6) = 1 - P(H_3)$

7 Conclusion

This paper demonstrates that neither the calculation nor the use of Bayes factors is straightforward. Bayes factors, as we have seen, are highly sensitive to the combination of the prior variance and the difference in the number of parameters in the rival models. We have also seen that Bayes factors do not provide a direct measure of support for one hypothesis over another, but rather, a measure of the change in the odds in favor of one hypothesis relative to another.

Bayes factors, then, do not completely overcome the problems with p-values, nor do they overcome the difficulty associated with the posterior odds ratio. Calculating Bayes factors requires specifying priors, which, by their very nature, reduces “objectivity.” Bayes factors are not posterior odds ratios, but neither do they provide direct measures of support. If we are interested in calculating what Bayes factors really tell us, we must live with the arbitrariness that accompanies that calculation.

References

- Albert, J. H. and S. Chib (1993, June). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association* 88(422), 669–679.
- Bartels, L. M. (1997, April). Specification uncertainty and model averaging. *American Journal of Political Science* 41(2), 641–674.
- Bartlett, M. (1957, December). A comment on d.v. lindley’s statistical paradox. *Biometrika* 44(3/4), 533–534.
- Berger, J. O. and T. Sellke (1987, March). Testing a point null hypothesis: The irreconcilability of p values and evidence. *Journal of the American Statistical Association* 82(397), 112–122.
- Casella, G. and E. I. George (1992, August). Explaining the gibbs sampler. *The American Statistician* 46(3), 167–174.
- Chib, S. (1995, December). Marginal likelihood from the gibbs output. *Journal of the American Statistical Association* 90(432), 1313–1321.
- Clarke, K. (2000, March). Testing nonnested models of international relations: Reevaluating realism.
- Drakos, N. (1995). Introduction to monte carlo methods. <http://csep1.phy.ornl.gov/mc/mc.html>.
- Earman, J. (1992). *Bayes or Bust? A Critical Examination of Bayesian Confirmation Theory*. Cambridge: The MIT Press.
- Edwards, A., H. Lindman, and L. Savage (1963). Bayesian statistical inference for psychological research. *Psychological Review* 70, 193–242.
- Gill, J. (1999, September). The insignificance of null hypothesis significance testing. *Political Research Quarterly* 52(3), 647–674.
- Han, C. and B. Carlin (2000, January). Mcmc methods for computing bayes factors: A comparative review.
- Huth, P., C. Gelpi, and D. S. Bennett (1993, September). The escalation of great power militarized disputes: Testing rational deterrence theory and structural realism. *American Political Science Review* 87(3), 609–623.
- Jackman, S. (2000, April). Estimation and inference via bayesian simulation. *American Journal of Political Science* 44(2), 375–404.

- Jeffreys, H. (1961). *Theory of Probability* (3 ed.). Oxford: Oxford University Press.
- Kass, R. and A. Raftery (1995, June). Bayes factors. *Journal of the American Statistical Association* 90(430), 773–795.
- Lavine, M. and M. J. Schervish (1999). Bayes factors: What they are and what they are not. *The American Statistician* 53(2), 119–122.
- Lehmann, E. L. (1986). *Testing Statistical Hypotheses* (2 ed.). New York: John Wiley.
- Raftery, A. E. (1994). Approximate bayes factors and accounting for model uncertainty in generalized linear models. Technical Report 255, Department of Statistics, University of Washington.
- Raftery, A. E. (1995). Bayesian model selection in social research (with discussion). *Sociological Methodology 1995 edited by P. V. Marsden*, Cambridge, MA: Blackwells.
- Reiter, D. and A. Stam (1998, June). Democracy, war initiation, and victory. *American Political Science Review* 92(2), 377–389.
- Salmon, W. (1990). Rationality and objectivity in science, or tom kuhn meets tom bayes. *Scientific Theories, Minnesota Studies in the Philosophy of Science 14*, edited by C.W. Savage, Minneapolis: University of Minnesota Press.
- Schervish, M. J. (1996, August). P values: What they are and what they are not. *The American Statistician* 50(3), 203–206.
- Smith, A. (1999, October). Testing theories of strategic choice: The example of crisis escalation. *American Journal of Political Science* 43(4), 1254–1283.
- Western, B. and S. Jackman (1994, June). Bayesian inference for comparative research. *American Political Science Review* 88(2), 412–423.

List of Tables

1	Marginal Likelihoods and Bayes factors for Chib data	22
2	Marginal Likelihoods and Bayes factors for Huth <i>et al.</i> data .	23
3	The Effect of Changing Variance on the Log of the Prior	24
4	The Effect of Changing Variance on the Log of the Prior when the Models are of Equal Size	25
5	Marginal Likelihoods and Bayes factors for Reiter and Stam data	26
6	Bayes Factors From the GLIB Software	27
7	Bayes Factors from the BIC Approximation	28

Table 1: Marginal Likelihoods and Bayes factors for Chib data

#	Prior Variance	Marginal Likelihoods ^a		Bayes factor
		Model <i>f</i>	Model <i>g</i>	
1	100	-30.94	-34.46	7.05
2	1000	-33.23	-36.75	7.05
3	10000	-35.53	-39.06	7.05

^aA positive Bayes factor is evidence in favor of model *f*

Table 2: Marginal Likelihoods and Bayes factors for Huth *et al.* data

#	Prior Variance	Marginal Likelihoods ^a		Bayes factor
		SR Model	RD Model	
1	100	-57.28	-53.06	-8.44
2	1000	-64.21	-64.46	0.51
3	10000	-71.06	-74.89	7.66

^aA positive Bayes factor is evidence in favor of the SR model.

Table 3: The Effect of Changing Variance on the Log of the Prior

#	Prior Variance	Log of the Prior	
		SR Model	RD Model
1	100	-19.35	-29.05
2	1000	-26.24	-40.51
3	10000	-33.15	-50.87

Table 4: The Effect of Changing Variance on the Log of the Prior when the Models are of Equal Size

#	Prior Variance	Log of the Prior	
		SR Model	RD Model
1	100	-19.35	-19.36
2	1000	-26.24	-26.24
3	10000	-33.15	-33.35

Table 5: Marginal Likelihoods and Bayes factors for Reiter and Stam data

#	Prior Variance	Marginal Likelihoods ^a		Bayes factor
		SR Model	RD Model	
1	100	-123.82	-79.72	-88.21
2	1000	-128.42	-90.19	-76.45
3	10000	-133.02	-101.60	-62.84

^aA positive Bayes factor is evidence in favor of the realist model.

Table 6: Bayes Factors From the GLIB Software

Reference Prior	Huth <i>et al.</i>	Reiter and Stam
$\varphi = 1$	-3.97	-78.07
$\varphi = 1.65$	-0.72	-73.69
$\varphi = 5$	6.09	-61.52

Table 7: Bayes Factors from the BIC Approximation

	Chib ^a	Reiter and Stam ^b	Huth <i>et al.</i> ^c
Bayes Factor	7.20	-73.47	-7.03
^a A positive Bayes factor is evidence in favor of model f . ^b A negative Bayes factor is evidence in favor of the realist model. ^c A negative Bayes factor is evidence in favor of the rational deterrence model.			