HYPOTHESIS ASSESSMENT USING THE BAYES FACTOR AND RELATIVE BELIEF RATIO

Z. Baskurt and M. Evans*

Department of Statistics, University of Toronto,
Toronto, Ontario M5S 3G3, Canada
**E-mail: mevans@utstat.utoronto.ca*
www.utstat.toronto.edu/mikevans/

The Bayes factor is commonly used for assessing the evidence for or against a given hypothesis $H_0 : \theta \in \Theta_0$, where Θ_0 is a subset of the parameter space. In this paper we discuss the Bayes factor and various issues associated with its use. A Bayes factor is seen to be intimately connected with a relative belief ratio which provides a somewhat simpler approach to assessing the evidence in favor of H_0. It is noted that, when there is a parameter of interest generating H_0, then a Bayes factor for H_0 can be defined as a limit and there is no need to introduce a discrete prior mass for Θ_0 or a prior within Θ_0. It is further noted that when a prior on Θ_0 does not correspond to a conditional prior induced by a parameter of interest generating H_0, then there is an inconsistency in prior assignments. This inconsistency can be avoided by choosing a parameter of interest that generates the hypothesis. A natural choice of a parameter of interest is given by a measure of distance of the model parameter from Θ_0. This leads to a Bayes factor for H_0 that is comparing the concentration of the posterior about Θ_0 with the concentration of the prior about Θ_0. The issue of calibrating a Bayes factor is also discussed and is seen to be equivalent to computing a posterior probability that measures the reliability of the evidence provided by the Bayes factor.

Keywords: Bayes factor; relative belief ratio; inequalities; method of concentration.

1. Introduction

Suppose we have the following ingredients available for a statistical problem: a statistical model $\{f_\theta : \theta \in \Theta\}$ given by a set of possible probability density functions with respect to a support measure μ on a sample space $\mathcal{X}$, a prior probability density π with respect to a support measure ν on the parameter space Θ, and the observed data $x \in \mathcal{X}$ which has been generated by one of the distributions in the model. Further suppose that our goal is to assess the hypothesis $H_0 : \theta \in \Theta_0$ where $\Theta_0 \subset \Theta$, namely, having observed x we want to assess the evidence that the true value of θ is in Θ_0.

A common way to approach this problem, based on the ingredients provided, is to compute the Bayes factor in favour of H_0. In fact we can only do this when $0 < \Pi(\Theta_0) < 1$ where Π is the prior probability measure. In this case the Bayes factor is given by

$$BF(H_0) = \frac{\Pi(\Theta_0 \,|\, x)}{1 - \Pi(\Theta_0 \,|\, x)} \Big/ \frac{\Pi(\Theta_0)}{1 - \Pi(\Theta_0)} \tag{1}$$

where $\Pi(\cdot \,|\, x)$ is the posterior probability measure. So $BF(H_0)$ measures the change from *a priori* to *a posteriori* in the odds in favour of H_0. If $BF(H_0) > 1$, then the data have lead to an increase in our beliefs that H_0 is true and so we have evidence in favour of H_0. If $BF(H_0) < 1$, then the data have lead to a decrease in our beliefs that H_0 is true and we have evidence against H_0.

Several questions and issues arise with the use of (1). First we ask why it is necessary to compare the odds in favour of H_0 rather than comparing the prior and posterior probabilities of H_0? Since this seems like a very natural way to make such a comparison we define the *relative belief ratio of* H_0 as

$$RB(H_0) = \frac{\Pi(\Theta_0 \,|\, x)}{\Pi(\Theta_0)} \tag{2}$$

whenever $0 < \Pi(\Theta_0)$. Also $RB(H_0) > 1$ is evidence in favour of H_0 while $RB(H_0) < 1$ is evidence against H_0. Note that, as opposed to the Bayes factor, the relative belief ratio is defined when $\Pi(\Theta_0) = 1$ but we then have that $RB(H_0) = 1$ for every x and this is uninteresting. The relationship between (1) and (2) is given by $BF(H_0) = (1-\Pi(\Theta_0))RB(H_0)/(1-\Pi(\Theta_0)RB(H_0))$ so $BF(H_0)$ and $RB(H_0)$ are 1-1 increasing functions of each other for fixed $\Pi(\Theta_0)$ but otherwise are measuring change in belief on different scales.

This raises the second question associated with both (1) and (2). In particular, what do we do when $\Pi(\Theta_0) = 0$ simply because Θ_0 is a lower dimensional subset of Θ? Certainly we want to assess hypotheses that correspond to lower dimensional subsets. The most common solution to this problem is to follow Jeffreys[12,13] and modify Π to be the mixture $\Pi_\gamma = \gamma\Pi_0 + (1-\gamma)\Pi$ where $\gamma \in (0, 1)$ and Π_0 is a probability measure concentrated on Θ_0. We then have that $\Pi_\gamma(\Theta_0) = \gamma$ and

$$BF(H_0) = m_0(x)/m(x) \tag{3}$$

where m_0 is the prior predictive density obtained from the model and Π_0, and m is the prior predictive density obtained from the model and Π.

Also, we have that $RB(H_0) = m_0(x)/(\gamma m_0(x) + (1-\gamma)m(x))$. While these calculations are formally correct, it is natural to ask if this approach is necessary as it does not seem reasonable that we should have to modify the prior Π simply because our hypothesis has $\Pi(\Theta_0) = 0$. In fact we will argue in Section 2 that this modification is often unnecessary as we can unambiguously define $BF(H_0)$ and $RB(H_0)$ by replacing Θ_0 on the right in (1) and (2) by a sequence of sets Θ_ϵ, where $\Pi(\Theta_\epsilon) > 0$ and $\Theta_\epsilon \downarrow \Theta_0$ as $\epsilon \downarrow 0$, and taking the limit. The limits obtained depend on how the sequence Θ_ϵ is chosen but this ambiguity disappears when H_0 is *generated* by a parameter of interest $\lambda = \Lambda(\theta)$ via $\Theta_0 = \Lambda^{-1}\{\lambda_0\}$ for some λ_0. In fact, when we have such a Λ then, using the definition via limits, $BF(H_0) = RB(H_0)$ and the two approaches to measuring change in belief are equivalent. Furthermore, if Π_0 is taken to be the conditional prior of θ given $\Lambda(\theta) = \lambda_0$, then $BF(H_0)$ defined as a limit equals (3) but this is not generally the case when Π_0 is chosen arbitrarily.

It is not always the case, however, that there is a parameter of interest generating H_0. In Section 3, we will argue that, in such a situation it is better that we choose such a Λ, as the introduction of γ and Π_0 can induce an inconsistency into the analysis. This inconsistency arises due to the fact that Π_0 does not necessarily arise from Π via conditioning. A natural choice is proposed in Section 3 where Λ is chosen so that $\Lambda(\theta)$ is a measure of the distance of θ from Θ_0. This is referred to as the *method of concentration* as the Bayes factor and relative belief ratio are now comparing the concentration of the posterior about Θ_0 with the concentration of the prior about Θ_0. If the data have lead to a greater concentration of the posterior distribution about Θ_0 than the prior, then this is naturally evidence in favour of H_0. This is dependent on the choice of the distance measure but now the conditional prior assignments on Θ_0 come from the prior Π.

A third issue is concerned with the calibration of $BF(H_0)$ or $RB(H_0)$, namely, when are these values large enough to provide convincing evidence in favour of H_0, or when are these values small enough to provide convincing evidence against H_0? In Section 4 we discuss some inequalities that hold for these quantities. For example, inequality (12) supports the interpretation that small values of $RB(H_0)$ provide evidence against H_0 while inequality (13) supports the interpretation that large values of $RB(H_0)$ provide evidence in favour of H_0. While these inequalities are *a priori,* the *a posteriori* probability (14) is a measure of the reliability of the evidence presented by $RB(H_0)$ given the specific data observed. In essence (14) is quantifying the uncertainty in the evidence presented by $RB(H_0)$.

For if (14) is small when $RB(H_0)$ is large, then there is a large posterior probability that the true value of the parameter of interest has an even larger relative belief ratio than the hypothesized value and so $RB(H_0)$ cannot be taken to be reliable evidence in favour of H_0 being true. On the other hand if (14) is large when $RB(H_0)$ is large, then indeed we can take the value $RB(H_0)$ as reliable evidence in favour of H_0. When $RB(H_0)$ is small, then a small value of (14) indicates that this is reliable evidence against H_0 and conversely for a large value of (14), although inequality (15) shows that this latter case is not possible. We also address the issue of when evidence against H_0 corresponds to practically meaningful evidence in the sense of whether or not we have detected a meaningful deviation from H_0. Similar comments apply to $BF(H_0)$.

The Bayes factor has been extensively discussed in the statistical literature. For example, Kass and Raftery[14] and Robert, Chopin, and Rousseau[16] provide excellent surveys. Our attention here is restricted to the case where the prior Π is proper. O'Hagan[15] defines a fractional Bayes factor and Berger and Perrichi[2] define an intrinsic Bayes factor for the case of improper priors.

Overall our purpose in this paper is to survey some recent results on the Bayes factor, provide some new insights into the meaning and significance of these results and illustrate their application through some simple examples. References to much more involved applications to problems of practical interest are also provided. Much of the technical detail is suppressed in this paper and can be found in Baskurt and Evans[1].

2. General Definition

Suppose that Π is discrete and that there is a parameter of interest $\lambda = \Lambda(\theta)$ generating H_0 via $\Theta_0 = \Lambda^{-1}\{\lambda_0\}$ where $0 < \Pi(\Theta_0) < 1$. We consider a simple example to illustrate ideas.

Example 1. *Discrete uniform prior on two points.*

Suppose that $\{f_\theta : \theta \in \Theta\}$ is a family of probability functions where we have $\Theta = \{1/4, 1/2\}^2, \theta = (\theta_1, \theta_2), f_{(\theta_1,\theta_2)}(x, y)$ is the Binomial$(m, \theta_1) \times$ Binomial(n, θ_2) probability function, Π is uniform on $\Theta, \lambda = \Lambda(\theta) = \theta_1 - \theta_2$ and we wish to assess the hypothesis $\Lambda(\theta) = \lambda_0 = 0$. So $\Theta_0 = \Lambda^{-1}\{\lambda_0\} = \{(1/4, 1/4), (1/2, 1/2)\} = \{(\omega, \omega) : \omega \in \{1/4, 1/2\}\}$ and $\Pi(\Theta_0) = 1/2$.

Suppose we observe $x = 2, y = 2, m = 4$ and $n = 5$. Then $\Pi(\Theta_0 \,|\, x, y) = 0.512, BF(H_0) = (0.512)(0.5)/[(1 - 0.512)(0.5)] = 1.049$ and $RB(H_0) = 2(0.512) = 1.024$. So both the Bayes factor and the relative belief ratio provide marginal evidence in favour of H_0 as beliefs have only increased slightly after having seen the data.

When Π_Λ is discrete we can write $RB(H_0) = \pi_\Lambda(\lambda_0 \,|\, x)/\pi_\Lambda(\lambda_0)$ and $BF(H_0) = \pi_\Lambda(\lambda_0 \,|\, x)(1 - \pi_\Lambda(\lambda_0))/[\pi_\Lambda(\lambda_0)(1 - \pi_\Lambda(\lambda_0 \,|\, x))]$ where π_Λ and $\pi_\Lambda(\cdot \,|\, x)$ denote the prior and posterior probability functions of Λ. Also note that, in the discrete case, the Bayes factor and relative belief ratio are invariant to the choice of Λ generating H_0 via $\Theta_0 = \Lambda^{-1}\{\lambda_0\}$. So, for example, we could take $\Lambda(\theta) = I_{\Theta_0}(\theta)$ where I_{Θ_0} is the indicator function of Θ_0. When there is a particular Λ of interest generating H_0 via $\Theta_0 = \Lambda^{-1}\{\lambda_0\}$, we will write $BF(\lambda_0)$ and $RB(\lambda_0)$ for the Bayes factor and relative belief ratios for H_0, respectively.

When $\Pi(\Theta_0) = 0$ a problem arises as $BF(H_0)$ and $RB(H_0)$ are not defined. As discussed in Baskurt and Evans[1], however, when there is a parameter of interest generating H_0 via $\Theta_0 = \Lambda^{-1}\{\lambda_0\}$, then sensible definitions are obtained via limits of sets shrinking to λ_0. For this we need to assume a bit more mathematical structure for the problem as described in Baskurt and Evans[1]. With this, the marginal prior density of $\lambda = \Lambda(\theta)$ is given by

$$\pi_\Lambda(\lambda) = \int_{\Lambda^{-1}\{\lambda\}} \pi(\theta) J_\Lambda(\theta)\, \nu_{\Lambda^{-1}\{\lambda\}}(d\theta) \tag{4}$$

with respect to volume (Lebesgue) measure ν_Λ on the range of Λ, where

$$J_\Lambda(\theta) = \left(\det(d\Lambda(\theta))(d\Lambda(\theta))^t\right)^{-1/2}, \tag{5}$$

$d\Lambda$ is the differential of Λ and $\nu_{\Lambda^{-1}\{\lambda\}}$ is volume measure on $\Lambda^{-1}\{\lambda\}$. Furthermore, the conditional prior density of θ given $\Lambda(\theta) = \lambda$ is

$$\pi(\theta \,|\, \lambda) = \pi(\theta) J_\Lambda(\theta)/\pi_\Lambda(\lambda) \tag{6}$$

with respect to $\nu_{\Lambda^{-1}\{\lambda\}}$. The posterior density of $\lambda = \Lambda(\theta)$ is then given by

$$\pi_\Lambda(\lambda \,|\, x) = \int_{\Lambda^{-1}\{\lambda\}} \pi(\theta) f_\theta(x) J_\Lambda(\theta)\, \nu_{\Lambda^{-1}\{\lambda\}}(d\theta)/m(x). \tag{7}$$

Note that in the discrete case the support measures are counting measure and in the continuous case these are things like length, area, volume and higher dimensional analogs.

Now suppose that $C_\epsilon(\lambda_0)$ is a sequence of neighborhoods shrinking nicely to λ_0 as $\epsilon \to 0$ with $\Pi_\Lambda(C_\epsilon(\lambda_0)) > 0$ for each ϵ; see Rudin[18] for the technical definition of 'shrinking nicely'. Then

$$RB(C_\epsilon(\lambda_0)) \to RB(\lambda_0) = \pi_\Lambda(\lambda_0 \,|\, x)/\pi_\Lambda(\lambda_0)$$

and

$$BF(C_\epsilon(\lambda_0)) \to BF(\lambda_0) = \pi_\Lambda(\lambda_0 \,|\, x)/\pi_\Lambda(\lambda_0)$$

as $\epsilon \to 0$ where π_Λ and $\pi_\Lambda(\cdot \,|\, x)$ are now the prior and posterior densities of Λ with respect to ν_Λ. Note that $BF(\lambda_0) = RB(\lambda_0)$ when $\Pi(\Theta_0) = 0$. So the Bayes factor and relative belief ratio of H_0 are naturally defined as limits. Note that the limiting relative belief ratio takes the same form in the discrete and continuous case but this is not true for the Bayes factor.

An alternative expression for the limiting relative belief ratio is shown in Baskurt and Evans[1] to be

$$RB(\lambda_0) = m(x \,|\, \lambda_0)/m(x) \tag{8}$$

where $m(\cdot \,|\, \lambda_0)$ is the conditional prior predictive density of the data given $\Lambda(\theta) = \lambda_0$. The equality

$$m(x \,|\, \lambda_0)/m(x) = \pi_\Lambda(\lambda_0 \,|\, x)/\pi_\Lambda(\lambda_0) \tag{9}$$

is the Savage-Dickey ratio result and this holds for discrete and continuous models; see Dickey and Lientz[4] and Dickey[5]. Note that (8) is not a Bayes factor in the discrete case. We conclude that, when H_0 arises from a parameter of interest via $\Theta_0 = \Lambda^{-1}\{\lambda_0\}$, there is no need to introduce a discrete mass of prior probability on Θ_0 to obtain the Bayes factor in favour of H_0 and the conditional prior on Θ_0 is unambiguously given by $\pi(\cdot \,|\, \lambda_0)$.

We consider a simple application of this.

Example 2. *Continuous prior with Λ specified.*

Suppose $\{f_\theta : \theta \in \Theta\}$ is the family of distributions where $\Theta = [0,1]^2, \theta = (\theta_1, \theta_2)\,, f_{(\theta_1,\theta_2)}(x,y)$ is the Binomial$(m, \theta_1) \times$ Binomial(n, θ_2) probability function, Π is uniform on $\Theta, \lambda = \Lambda(\theta) = \theta_1 - \theta_2$ and we wish to assess the hypothesis $\lambda = \lambda_0 = 0$. So $\Theta_0 = \Lambda^{-1}\{\lambda_0\} = \{(\theta,\theta) : \theta \in [0,1]\}$ and $\Pi(\Theta_0) = 0$. Using (5) we have $J_\Lambda(\theta) = 1/\sqrt{2}$ and, since $\nu_{\Lambda^{-1}\{\lambda\}}$ is length measure on $\Lambda^{-1}\{\lambda\}$, applying this to (4) gives $\pi_\Lambda(\lambda) = 1 - \lambda$ when $\lambda \geq 0, \pi_\Lambda(\lambda) = 1 + \lambda$ when $\lambda \leq 0$ and $\pi_\Lambda(\lambda) = 0$ otherwise. To compute $RB(\lambda) = \pi_\Lambda(\lambda \,|\, x, y)/\pi_\Lambda(\lambda)$ for a general Λ, we will typically have to sample from the prior to obtain π_Λ but here we have an exact expression for the prior. From (6) we see that the conditional prior density of θ given $\lambda = \Lambda(\theta)$, with respect to length measure on $\{\theta : \lambda = \Lambda(\theta)\}$, is given by $\pi(\theta \,|\, \lambda) = 1/\sqrt{2}(1-\lambda)$ when $\lambda \geq 0$ and by $\pi(\theta \,|\, \lambda) = 1/\sqrt{2}(1+\lambda)$ when $\lambda \leq 0$. Therefore, the conditional prior is uniform.

The posterior distribution of θ is given by

$$(\theta_1, \theta_2) \mid (x, y) \sim \text{Beta}\,(x+1, m-x+1) \times \text{Beta}\,(y+1, n-y+1)\,.$$

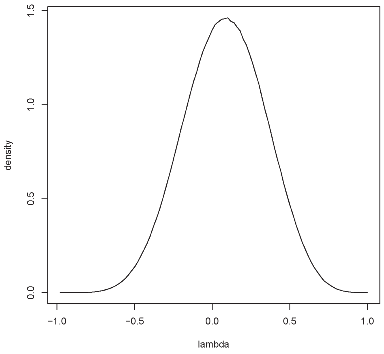

Fig. 1. Plot of estimate of the posterior density of $\lambda = \Lambda(\theta)$ in Example 2.

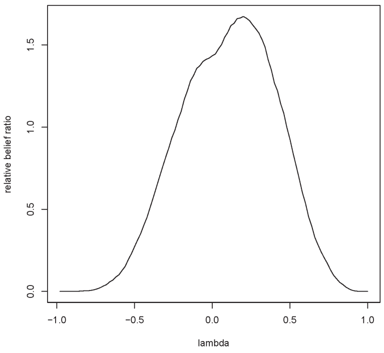

Fig. 2. Plot of estimate of the relative belief ratios of $\lambda = \Lambda(\theta)$ in Example 2.

Suppose we observe $x = 2, y = 2, m = 4$ and $n = 5$. In Figure 1 we present a plot of the posterior density of $\lambda = \Lambda(\theta)$ and in Figure 2 a plot of the relative belief ratio $RB(\lambda)$ as a function of λ based on samples of 10^5 from the prior and posterior. The maximum value of $RB(\lambda)$ is 1.68 and this occurs at $\hat{\lambda} = 0.22$, so $RB(\hat{\lambda}) = 1.68$. The relative belief ratio in favour of H_0 is given by $RB(0) = 1.47$ and so we have evidence in favour of H_0. This evidence doesn't seem overwhelming, although it does say that the data has lead to an approximate 50% increase in the "probability" of Θ_0.

3. The Method of Concentration

It is a feature of many problems, however, that H_0 does not arise from an obvious parameter of interest $\lambda = \Lambda(\theta)$ and so appropriate definitions of $BF(H_0)$ and $RB(H_0)$ are ambiguous when $\Pi(\Theta_0) = 0$.

Example 3. *Continuous prior with Λ unspecified.*

Suppose that the situation is identical to that described in Example 2 but now we do not declare that we want to make inference about $\lambda = \Lambda(\theta) = \theta_1 - \theta_2$. Rather it is simply specified that we wish to assess the hypothesis $\Theta_0 = \{(\omega, \omega) : \omega \in [0, 1]\}$, namely, we only specify that we are interested in determining whether or not $\theta_1 = \theta_2$. Of course, $\Theta_0 = \Lambda^{-1}\{0\}$ but there are many such Λ generating H_0 and it is not clear, given the statement of the problem, which we should use. It seems clear from (4) and (5), however, that $RB(H_0) = RB(\lambda_0)$ will depend on the Λ we use to obtain the relative belief ratio.

It is common practice in such situations to follow Jeffreys and replace Π by the prior $\Pi_\gamma = \gamma\Pi_0 + (1-\gamma)\Pi$ to compute $BF(H_0)$ and $RB(H_0)$. From (3) we see that, when Π_0 corresponds to the conditional prior of $\lambda = \Lambda(\theta)$ given that $\lambda = \lambda_0$, then $m_0(x) = m(x \,|\, \lambda_0)$ and (3) equals (8). Again there is no need to introduce γ.

On the other hand, if Π_0 is not equal to the conditional prior based on Π for *some* Λ, then a fundamental inconsistency arises in the Bayesian analysis, as the conditional beliefs on Θ_0 do not arise from the prior Π. The existence of such an inconsistency when Π is discrete is unacceptable and there is no reason to allow this in the continuous case either.

Example 4. *Discrete uniform prior on many points.*

Consider the context of Example 1 where now $\Theta = \{1/k, 2/k, \ldots, (k-2)/(k-1)\}^2$ for some positive integer k and Π is the uniform prior on this set. Suppose again we want to assess the hypothesis $\Theta_0 = \{(\omega, \omega) : \omega \in 1/k, 2/k, \ldots, (k-2)/(k-1)\}\}$. The hypothesis being assessed is clearly the assertion that $\theta_1 = \theta_2$. Note that in this case, however we choose Λ generating H_0, the relative belief ratio in favour of H_0 is the same and the conditional prior given Θ_0 is unambiguously the uniform prior on Θ_0. Furthermore, when k is very large we can think of the continuous problem in Example 3 as an approximation to this problem. Without specifying a Λ in Example 3, however, we are not specifying how this approximation is taking place.

Avoiding the inconsistency requires the choice of a suitable Λ that generates H_0 via $\Theta_0 = \Lambda^{-1}\{\lambda_0\}$ and then using the relative belief ratio $RB(\lambda_0)$. Faced with the option of either choosing γ and Π_0 or choosing a Λ that generates H_0, the latter seems preferable as it ensures that beliefs are being assigned consistently in the problem.

The effect of the inconsistency can be seen directly via a generalization of the Savage-Dickey ratio result (9) due to Verdinelli and Wasserman[19] and that is also discussed in Baskurt and Evans[1]. This result was derived to aid in the computation of (3) but in fact it relates the Bayes factor obtained using the Jeffreys approach via γ and Π_0 and that obtained via the definition in Section 2 as a limit. This can be written in two ways as

$$\begin{aligned} m_0(x)/m(x) &= RB(\lambda_0)E_{\Pi_0}\left(\pi(\theta\,|\,\lambda_0,x)/\pi(\theta\,|\,\lambda_0)\right) \\ &= RB(\lambda_0)E_{\Pi(\cdot\,|\,\lambda_0,x)}\left(\pi_0(\theta)/\pi(\theta\,|\,\lambda_0)\right). \qquad (10) \end{aligned}$$

The first equality in (10) says that (3) equals the Bayes factor obtained as a limit in Section 2, times the expected conditional relative belief ratio of θ computed as a limit via $\Pi(\cdot\,|\,\lambda_0)$, where the expectation is taken with respect to the prior Π_0 placed on Θ_0. It is interesting to note that the adjustment factor involves relative belief ratios. The second equality in (10) says that (3) equals the Bayes factor obtained as a limit in Section 2, times the expected ratio of the prior π_0 evaluated at θ to the conditional prior induced by Π evaluated at θ, given that $\lambda_0 = \Lambda(\theta)$, where the expectation is taken with respect to the conditional posterior induced by the prior Π, given that $\lambda_0 = \Lambda(\theta)$. So if π_0 is very different from any conditional prior $\pi(\cdot\,|\,\lambda_0)$ induced by Π, then we can expect a big differences in the Bayes factors obtained by the two approaches.

To avoid this inconsistency a natural choice of Λ in such a problem is to take $\Lambda = d_{\Theta_0}$, where $d_{\Theta_0}(\theta)$ is a measure of the distance θ is from Θ_0. Therefore, $d_{\Theta_0}(\theta) = 0$ if and only if $\theta \in \Theta_0$ and $\Theta_0 = \Lambda^{-1}\{0\}$. With this choice the Bayes factor (and relative belief ratio) $RB(0)$ is a comparison of the concentration of the posterior distribution about Θ_0 with the concentration of the prior distribution about Θ_0. If $RB(0) > 1$, then the posterior distribution is concentrating more about Θ_0 than the prior and we have evidence in favour of H_0. This seems quite natural as when H_0 is true we expect the posterior distribution to assign more of its mass near Θ_0 than the prior, otherwise the data is providing evidence against H_0. Under weak conditions $RB(0)$ will converge to 0 when H_0 is false and to ∞ when H_0 is true, as the amount of data increases.

Of course, there is still an arbitrariness as there are many possible choices of distance measure. But this arbitrariness is in essence an unavoidable consequence of a problem that is not fully specified. The problem is similar to the Borel paradox in probability theory where, in general, there is no unique conditional probability measure associated with conditioning on a set of measure 0 even though this may make perfect sense as in Example 2. The way out of this is to specify a function that generates the set but the conditional distribution depends on the function chosen. Our recommendation is that we should address the problem in a way that guarantees beliefs assigned in a consistent way and this effectively entails choosing a Λ that generates H_0. If we choose Λ in an intuitively satisfying way then this adds greater support for this approach. Setting $\Lambda = d_{\Theta_0}$ for some distance measure d_{Θ_0} satisfies these criteria.

In a number of problems, see the discussion in Section 5, we have chosen d_{Θ_0} to be squared Euclidean distance so we are effectively using least squares. This choice often exhibits a very nice property as expressed in the following result.

Proposition 1. $RB(\lambda_0)$ is the same for all Λ in the set $\{\Lambda : J_\Lambda(\theta)$ is constant and nonzero for all $\theta \in \Lambda^{-1}\{\lambda_0\}\}$.
Proof: From (4) and (7) we have that

$$\begin{aligned} RB(\lambda_0) &= \frac{\pi_\Lambda(\lambda_0 \,|\, x)}{\pi_\Lambda(\lambda_0)} = \frac{\int_{\Lambda^{-1}\{\lambda_0\}} \pi(\theta) f_\theta(x) J_\Lambda(\theta)\, \nu_{\Lambda^{-1}\{\lambda_0\}}(d\theta)/m(x)}{\int_{\Lambda^{-1}\{\lambda_0\}} \pi(\theta) J_\Lambda(\theta)\, \nu_{\Lambda^{-1}\{\lambda_0\}}(d\theta)} \\ &= \frac{\int_{\Theta_0} \pi(\theta) f_\theta(x)\, \nu_{\Theta_0}(d\theta)/m(x)}{\int_{\Theta_0} \pi(\theta)\, \nu_{\Theta_0}(d\theta)} \end{aligned}$$

where the last equality follows from $\Theta_0 = \Lambda^{-1}\{\lambda_0\}$ and $\nu_{\Lambda^{-1}\{\lambda_0\}} = \nu_{\Theta_0}$ since this measure is determined by the geometry of Θ_0 alone. □

As already noted, Proposition 1 always holds when Π is discrete because the parameter space is countable and with the discrete topology on Θ, all functions are continuously differentiable with $J_\Lambda(\theta) \equiv 1$. Also, whenever $\Theta_0 = \{\theta_0\}$ for some $\theta_0 \in \Theta$, then any Λ continuously differentiable at θ_0 that generates $\Theta_0 = \{\theta_0\}$ clearly satisfies the condition of Proposition 1.

When $\Theta_0 = \mathcal{L}(A)$ for some $A \in R^{k\times m}$ of rank m and $\Lambda(\theta) = d_{\Theta_0}(\theta) = ||\theta - (A'A)^{-1}A'\theta||^2$ is the squared Euclidean distance of θ from Θ_0, then $J_\Lambda(\theta) = ||\theta - (A'A)^{-1}A'\theta||^{-1}/2 = \Lambda^{-1/2}(\theta)/2$ and so if $\lambda_0 > 0$ the conditions of Proposition 1 are satisfied. When $\lambda_0 \to 0$ we have that

$$RB(\lambda_0) \to RB(0) = \frac{\int_{\Theta_0} \pi(\theta) f_\theta(x)\, \nu_{\Theta_0}(d\theta)/m(x)}{\int_{\Theta_0} \pi(\theta)\, \nu_{\Theta_0}(d\theta)} \tag{11}$$

which is independent of Λ. So when Λ generating H_0 satisfies the requirement of Proposition 1, we see that volume distortions induced by Λ, as measured by $J_\Lambda(\theta)$, do not affect the value of the relative belief ratio.

We consider an example using squared Euclidean distance.

Example 5. *Continuous prior using concentration.*

Suppose the situation is as in Example 3. We take $\Lambda_*(\theta) = d_{\Theta_0}(\theta)$ to be the squared Euclidean distance of θ from $\Theta_0 = \{(\omega, \omega) : \omega \in [0,1]\}$. It is clear that $\Lambda_*(\theta) = (\theta_1 - \theta_2)^2/2$, $J_{\Lambda_*}(\theta) = \Lambda_*^{-1/2}(\theta)/2$ and (11) applies. Now notice that the Λ used in Example 2 also satisfies the conditions of Proposition 1 and so $\lim_{\lambda_0 \to 0} RB(\lambda_0) = RB(0)$ must be the same as what we get when using Λ_* and (11). Therefore, for the data as recorded in Example 2 we also have $RB(0) = 1.47$ when using the method of concentration.

The outcome in Example 5 is characteristic of many situations when taking $d_{\Theta_0}(\theta)$ to be the squared Euclidean distance of θ from Θ_0.

4. Calibration

In Baskurt and Evans[1] several *a priori* inequalities were derived that support the interpretations of $BF(\lambda_0)$ or $RB(\lambda_0)$ as evidence for or against H_0. These are generalizations to the Bayesian context of inequalities derived in Royall[17] for likelihood inferences. For example, it can be proved that

$$M\left(m(X \,|\, \lambda_0)/m(X) \le RB(\lambda_0) \,|\, \lambda_0\right) \le RB(\lambda_0) \tag{12}$$

and

$$M(\cdot \,|\, \lambda_0) \times \Pi_\Lambda \left(m(X \,|\, \lambda)/m(X) \ge RB(\lambda_0)\right) \le \left(RB(\lambda_0)\right)^{-1}. \tag{13}$$

In both inequalities we consider $RB(\lambda_0)$ as a fixed observed value of the relative belief ratio and $X \sim M\left(\cdot \mid \lambda_0\right)$ where $M\left(\cdot \mid \lambda_0\right)$ is the conditional prior predictive measure given that $\Lambda(\theta) = \lambda_0$. So inequality (12) says that the conditional prior probability, given that H_0 is true, of obtaining a value of the relative belief ratio (recall (9)) of H_0 smaller than the observed value is bounded above by $RB(\lambda_0)$. So if $RB(\lambda_0)$ is very small this probability is also very small and we can consider a small value of $RB(\lambda_0)$ as evidence against H_0. In (13) $\lambda \sim \Pi_\Lambda$ independent of X and the inequality says that the conditional prior probability, given that H_0 is true, of obtaining a larger value of the relative belief ratio at a value λ generated independently from Π_Λ, is bounded above by $\left(RB(\lambda_0)\right)^{-1}$. So if $RB(\lambda_0)$ is very large then, when H_0 is true, it is extremely unlikely that we would obtain a larger value of $RB(\lambda)$ at a value λ that is *a priori* reasonable. So we can consider large values of $RB(\lambda_0)$ as evidence in favour of H_0. Note that these inequalities also apply to the Bayes factor in the continuous case and similar inequalities can be derived for the Bayes factor in the discrete case.

We now consider the reliability or the uncertainty in the evidence given by $RB(\lambda_0)$. We will measure the reliability of this evidence by comparing it to the evidence in favour of alternative values of λ. For, if $RB(\lambda_0)$ is very large, so we have strong evidence in favour of H_0, but $RB(\lambda)$ is even larger for values of $\lambda \neq \lambda_0$, then this casts doubt on the reliability of the evidence in favour of H_0. The probabilities in (12) and (13) are *a priori* measures of the reliability of the evidence given by $RB(\lambda_0)$. For inequality (12) tells us that, when $RB(\lambda_0)$ is very small, there is little prior probability of getting an even smaller value of this quantity when H_0 is true. Similarly, (13) tells us that, when $RB(\lambda_0)$ is very large, there is little prior probability of getting an even larger value for $RB(\lambda)$ for some λ when H_0 is true. Based upon fundamental considerations, however, we know that we need to measure the reliability using posterior probabilities. Accordingly, we propose to use the posterior tail probability

$$\Pi\left(RB(\Lambda(\theta)) \leq RB(\lambda_0)\,|\,x\right) \tag{14}$$

to measure the reliability, or equivalently quantify the uncertainty, in the evidence given by $RB(\lambda_0)$. We see that (14) is the posterior probability of a relative belief ratio (and Bayes factor in the continuous case) $RB(\lambda)$ being no larger than $RB(\lambda_0)$. If $RB(\lambda_0)$ is large and (14) is small (see Baskurt and Evans[1] for examples where this occurs), then evidence in favour of H_0, as expressed via $RB(\lambda_0)$, needs to be qualified by the fact that our posterior beliefs are pointing to values of λ where the data have lead to an even bigger increase in belief. In such a situation the evidence in favour of H_0 does not seem very reliable. If $RB(\lambda_0)$ is large and (14) is also large, then we have reliable evidence in favour of H_0. Similarly, if $RB(\lambda_0)$ is small and (14) is small, then we have reliable evidence against H_0, while a large value of (14) suggests the evidence against H_0, as expressed through $RB(\lambda_0)$, is not very reliable. In fact, the *a posteriori* inequality

$$\Pi\left(RB(\Lambda(\theta)) \leq RB(\lambda_0)\,|\,x\right) \leq RB(\lambda_0) \tag{15}$$

is established in Baskurt and Evans[1]. Inequality (15) shows that a very small value of $RB(\lambda_0)$ is always reliable evidence against H_0.

We note that (14) is not interpreted like a P-value. The evidence for or against H_0 is expressed via $RB(\lambda_0)$ and (14) is a measure of the reliability of this evidence. A similar inequality exists for the Bayes factor in the discrete case.

It is a substantial advantage of the use of the Bayes factor and the relative belief ratio that they can express evidence both for and against hypotheses. A weakness of this approach is that somewhat arbitrary scales have been created for comparison purposes for the Bayes factor. For example, according to such a scale a Bayes factor of 20 is considerable evidence in favour of λ_0. But, as shown in Baskurt and Evans[1], it is possible that there are other values of λ for which the Bayes factor is even larger and this leads one to doubt the evidence in favour of λ_0. If, however, such values have only a small amount of posterior weight, then the evidence in favour of λ_0 seems more compelling. Assessing this is the role of (14).

As with the use of P-values, however, we need to also take into account the concept of practical significance when assessing hypotheses. It is clear that $RB(\lambda_0)$, even together with (14), doesn't do this. For example, suppose $RB(\lambda_0)$ is small and (14) is small so we have evidence against H_0, or suppose $RB(\lambda_0)$ is large while (14) is small or at least not large. In such a situation it is natural to compute $(\hat{\lambda}, RB(\hat{\lambda}))$ where $\hat{\lambda} = \arg\sup RB(\lambda)$ as $\hat{\lambda}$ is in a sense the value of λ best supported by the data. If the difference between $\hat{\lambda}$ and λ_0 is not of practical significance, then it seems reasonable to proceed as if H_0 is true. So the approach discussed here can also take into account the notion of practical significance.

We consider an example involving computing (14).

Example 6. *Measuring the reliability of a Bayes factor.*

Consider the context discussed in Example 2. We had $\lambda = \Lambda(\theta) = \theta_1 - \theta_2$ and we wished to assess the hypothesis $\Lambda(\theta) = \lambda_0 = 0$, so $\Theta_0 = \Lambda^{-1}\{\lambda_0\} = \{(\theta, \theta) : \theta \in [0, 1]\}$. For the data specified there, we obtained $RB(0) = 1.47$ and we have evidence in favour of H_0 although not overwhelmingly strong. Using simulation we computed (14) as equal to 0.42. This indicates that the posterior probability of obtaining a larger value of the Bayes factor is 0.58 so the evidence in favour of H_0 is not particularly reliable. Of course we have very small samples here so this is not surprising.

The maximum value of $RB(\lambda)$ is 1.68 and this occurs at $\hat{\lambda} = 0.22$, so $RB(\hat{\lambda}) = 1.68$. If in the particular application we also thought that $\hat{\lambda} = 0.22$ does not differ meaningfully from 0, then it seems reasonable to proceed as if H_0 is true.

5. Conclusions

This paper has shown that is possible to provide a sensible definition of a Bayes factor in favour of a hypothesis $H_0 : \theta \in \Theta_0$, in contexts where Θ_0 has prior probability 0 simply because it is a lower dimensional subset of Θ and the prior Π is continuous on Θ. This is accomplished without the need to introduce the mixture prior $\Pi_\gamma = \gamma\Pi_0 + (1 - \gamma)\Pi$ which requires the specification of γ and Π_0, in addition to Π. Our approach avoids the need for the discrete mass γ but more importantly, when Π_0 is not given by the conditional prior of θ given that $\Lambda(\theta) = \lambda_0$ for some Λ generating H_0, it avoids an inconsistency in the assignment of prior beliefs. This provides a unified approach to Bayesian inference as we no longer have to treat estimation and hypothesis assessment problems separately, namely, we do not have to modify an elicited prior Π just to deal with a hypothesis assessment. In situations where H_0 is not generated by a parameter of interest, then the method of concentration is seen to be a natural approach to choosing such a Λ. We have also discussed the calibration of a Bayes factor or relative belief ratio and have shown that this is intimately connected with measuring the reliability of the evidence presented by a Bayes factor or relative belief ratio. Finally we have shown how a maximized Bayes factor or relative belief ratio can be used to assess the practical significance of the evidence presented by these quantities. While we have illustrated these ideas via simple examples in this paper much more substantive and practical applications can be found in Evans, Gilula and Guttman[6] and Evans, Gilula, Guttman and Swartz,[8] Cao, Evans and Guttman,[3] Evans, Gilula and Guttman[7] and Baskurt and Evans.[1]

Other inferences are closely related to those discussed in this paper. For example when we have a parameter of interest Λ, then the value $\hat{\lambda}$ maximizing $RB(\lambda)$, referred to as the *least relative surprise estimator (LRSE)*, can be used to estimate λ. It is shown in Evans and Jang[10] that $\hat{\lambda}$ is either a Bayes rule or a limit of Bayes rules where the losses are derived from the prior. A γ-credible region for λ is given by $C_\gamma(x) = \{\lambda_0 : \Pi\left(RB(\Lambda(\theta)) \leq RB(\lambda_0) \,|\, x\right) > 1 - \gamma\}$ and is referred to as a *γ-relative surprise region.* A variety of optimality properties have been established for these regions, when compared to other rules for the construction of Bayesian credible regions, in Evans, Guttman and Swartz[9], and Evans and Shakhatreh.[11]

References

1. Baskurt and Evans (2011) Inequalities for Bayes factors and relative belief ratios. Tech. Rep. 1105, Dept. of Stat., U. of Toronto.
2. Berger, J.O. and Perrichi, R.L. (1996). The intrinsic Bayes factor for model selection and prediction. J. Amer. Stat. Assoc., 91, 10-122.
3. Cao, Y., Evans, M. and Guttman, I. (2010). Bayesian factor analysis via concentration. Tech. Rep. No. 1003, Dept. of Stat., U. of Toronto.
4. Dickey, J.M. and Lientz, B.P. (1970). The weighted likelihood ratio, sharp hypotheses about chances, the order of a Markov chain. Ann. Math. Stat., 41, 1, 214-226.

5. Dickey, J.M. (1971). The weighted likelihood ratio, linear hypotheses on normal location parameters. Ann. Stat., 42, 204-223.
6. Evans, M., Gilula, Z., and Guttman, I. (1993) Computational issues in the Bayesian analysis of categorical data : loglinear and Goodman's RC model. Stat. Sin., 3, 391-406.
7. Evans, M., Gilula, Z., and Guttman, I. (2012) Conversion of ordinal attitudinal scales: an inferential Bayesian approach. To appear in Quant. Mark. and Eco.
8. Evans, M., Gilula, Z., Guttman, I., and Swartz, T. (1997) Bayesian analysis of stochastically ordered distributions of categorical variables. J. Amer. Stat. Assoc., 92, 437, 208-214.
9. Evans, M., Guttman, I. and Swartz, T. (2006) Optimality and computations for relative surprise inferences. Can. J. of Stat., 34, 1, 113-129.
10. Evans, M. and Jang, G-H. (2011) Inferences from prior-based loss functions. Tech. Rep. 1104, Dept. of Stat, U. of Toronto.
11. Evans, M. and Shakhatreh, M. (2008) Optimal properties of some Bayesian inferences. Elec. J. of Stat., 2, 1268-1280.
12. Jeffreys, H. (1935) Some tests of significance, treated by the theory of probability. Proc. Camb. Phil. Soc., 31, 203- 222.
13. Jeffreys, H. (1961) *Theory of Probability,* 3rd edn. (Oxford University Press, Oxford, U.K.)
14. Kass, R.E. and Raftery, A.E. (1995) Bayes factors. J. Amer. Stat. Assoc., 90, 430, 773-795.
15. O'Hagan, A. (1995) Fractional Bayes factors for model comparisons (with discussion). J. Royal Stat. Soc. B, 56, 3-48.
16. Robert, C.P., Chopin, N. and Rousseau, J. (2009) Harold Jeffreys's Theory of Probability Revisited (with discussion). Stat. Sci., 24, 2, 141-172.
17. Royall, R. (2000) On the probability of observing misleading statistical evidence (with discussion). J. Amer. Stat. Assoc., 95, 451, 760-780.
18. Rudin, W. (1974) *Real and Complex Analysis*, 2nd edn. (McGraw-Hill, New York).
19. Verdinelli, I. and Wasserman, L. (1995) Computing Bayes factors using a generalization of the Savage-Dickey density ratio. J. Amer. Stat. Assoc., 90, 430, 614-618.