Corrected ABF

Karl Tayeb

Bayes factors

1. Given data $\mathcal{D}$ and two hypotheses H_1 and H_2, the Bayes factor (BF) BF_{12} quantifies the evidence in favor of hypothesis H_1 against hypothesis H_2. The BF is defined as the ratio of posterior odds to prior odds (note the BF is the same regardless of value of the prior odds, observing data $\mathcal{D}$ provides the same evidence for each model, regardless of the prior specificiation).

$$\text{posterior odds} = BF_{12} \times \text{prior odds}$$

2. The BF can be computed as the ratio of marginal likelihoods under each model. *Marginal* because specifying a hypothesis H involves specifying a prior for, and integrating over, any unobserved parameters. When the hypothesis are *simple* the BF is simply a likelihood ratio. We are interested particularly in the case of a simple null.

$$BF_{12} = \frac{P(\mathcal{D}|H_1)}{P(\mathcal{D}|H_2)}$$

3. When there are unkown parameters in the model, evaluating the marginal likelihood requires evaluating an integral. In special cases (e.g. exponential family with conjugate prior) the marginal likleihoods can be computed analytically. However, in general there may be no analytic solution. In high dimensional settings the integration can quickly become intractable, so there is considerable interest in developing tractable approximations to the marginal likleihood.

$$P(\mathcal{D}|H_i) = \int P(\mathcal{D}|\theta_i, H_i) P(\theta_i|H_i) d\theta_i$$

Laplace approximation

1. The Laplace approximation is a general technique for approximating integrals of nonnegative functions. Suppose $f : \mathbb{R} \to \mathbb{R}^+$ is a nonegative function, and we want to approximate the integral

$$I = \int f(x) dx,$$

We can define $h(x) = \log f(x)$ and approximate h with a second order Taylor expansion around x^*, $\hat{h}(x) = h(x^*) + h^{'}(x^*)(x - x^*) + \frac{1}{2}h^{''}(x^*)(x - x^*)^2$. Given certain regularity conditions, if x^* is a (unique) maximizer of h then the first order term vanished and $h^{''}(x^*) < 0$. Substituting the approximation $f(x) \approx \exp \hat{h}(x))$ we can approximate I with a Gaussian integral

$$\hat{I} = \exp \hat{h}(x^*) \left(-\frac{2\pi}{h^{''}(x^*)} \right)^{1/2}$$

2. The Laplace approximation requires knowledge of the MAP to form the approximation to $\log P(\mathcal{D}, \theta_i | H_i)$ which necessitate specification of the prior. It is often convenient to omit the prior and approximate only $\hat{h}(\theta_i) \approx \log P(\mathcal{D}|\theta_i, H_i)$.

$$\hat{I}_{MLE} = \int \exp\{\hat{h}(\theta_i)\} P(\theta_i | H_i).$$

Here, $\hat{h}$ is a Taylor expansion around the MLE. With a normal prior on θ_i, the integral can be computed analytically. This approximation is attractive because it partitions the analysis– we can summarize the information that the data provide on the parameter of interest, and combine in with the prior later. Except in cases where the data contain very little information on θ_i relative to the prior, this approximation will also be good. The BF may be approximated by approximating the marginal likelihood in both the numerator and the denominator seperately. As noted in Kass and Raftery the Laplace approximation achieved relative error of $O(n^{-1})$, and so does the approximation of the ratio. Both the Laplace approximation and this variant are advocated by Kass and Raftery in their work popularizing BFs.

3. These approximation can be characterized as "local"; they are exact at a point, and approximation quality decays as we move away from that point. The Laplace approximation makes the choice of making the approximation exact at the maximum of f. This is a good choice for integrtion of a non-negative function because this neighborhood contributes a lot of mass to the integral. Intuitively, the Laplace approximation will be best if f is strongly peaked about it's maximum, so that the region of the parameter space that contribute most to the integral are well approximated.
4. In the case of a simple null hypothesis (with the null in the denominator), we can also think of the BF as integrating over the likelihood ratio:

$$BF_{10} = \int \frac{P(\mathcal{D}|\theta_1, H_1)}{P(\mathcal{D}|H_0)} P(\theta_1 | H_1) d\theta_1.$$

We find this perspective instructive for the forthcoming comparsion of Wakefield's approximate Bayes Factor (ABF) and the alternative Laplace-type approximation we advocate for here. As we will see, the ABF uses the same local approximation for both the numerator and the denominator of the likelihood ratio. However, in the case we are interested in there is no need for approximation of the denominator. Approximating the numerator only, or equivalently, approximating the likelihood ratio directly can dramatically improve the approximation when the approximation of the denominator is poor.

Asymptotic approximation and the approximate Bayes Factor

1. Asymptotically, $\hat{\beta} \sim N(\beta, s^2)$. Where $\hat{\beta}$ is the MLE and s is it's standard error. Since $\hat{\beta}$ is approximately sufficient (for β), that is $P(\mathcal{D}|\hat{\beta}, \beta) \approx P(\mathcal{D}|\hat{\beta})$ we can approximate the likelihood ratio
$$\frac{N(\hat{\beta}; \beta, s^2)}{N(\hat{\beta}; 0, s^2)},$$
Integrating over a normal prior $\beta \sim N(0, \sigma^2)$ gives the approximate Bayes Factor introduced by Jon Wakefield
$$ABF = \frac{N(\hat{\beta}; 0, s^2 + \sigma^2)}{N(\hat{\beta}; 0, s^2)}.$$
2. This approximation is appealing because it only depends on the MLE for β and it's standard error. Focussing on the approximation of the likelihood ratio, we see that the likelihood ratio approximation is equal to 1 and $\beta = 0$. Wakefield notes that the ABF is consistent in the sense that the ABF will lead to the correct *decision* asymptotically. That is, as $n \to \infty$ the ABF will grow unbounded under the alternative, and shrink to 0 under the null. In this report, we want to highlight that the ABF may not provide a good approximation of the BF itself, *especially* when evidence against the null is strong.
3. The ABF uses an asymptotic approximation in both the numerator *and the denominator* of the likelihood ratio. While the asymptotic approximation is good near the MLE, it can be quite bad in the tail. If the z score $z = \hat{\beta}/s$ is large, the ABF can be using a very poor estimate of the null likelihood– that is $N(\hat{\beta}; 0, s^2)$ is a poor approximation of $P(\hat{\beta}|\beta = 0)$. Put another way, the ABF can be thought of as using a 2nd order Taylor approximation of the log-likelihood about the MLE for both the numerator and denominator. This choice of approximation may be good for the numerator– but if $\hat{\beta}$ is "far" from 0, it may by a very poor approximation for the denominator, the likelihood under the null model.
4. In many cases the likelihood under the null model is not hard to compute, and is often reported by standard statistical software. In applications where having accurate approximation of the BFs is important (e.g. so that they are comparable across variables, as in the variable selection example presented below) it may be prudent to skip approximating the denominator of the likelihood ratio.

Laplace approximate Bayes Factor (LABF)

1. Observing that we do not need to approximate the denominator of the BF, an alternative is to use the Lapalce approximation only to approximate the marginal likelihood in the numerator. Equivalently, we can think of a Laplace approximation to the likelihood ratio.

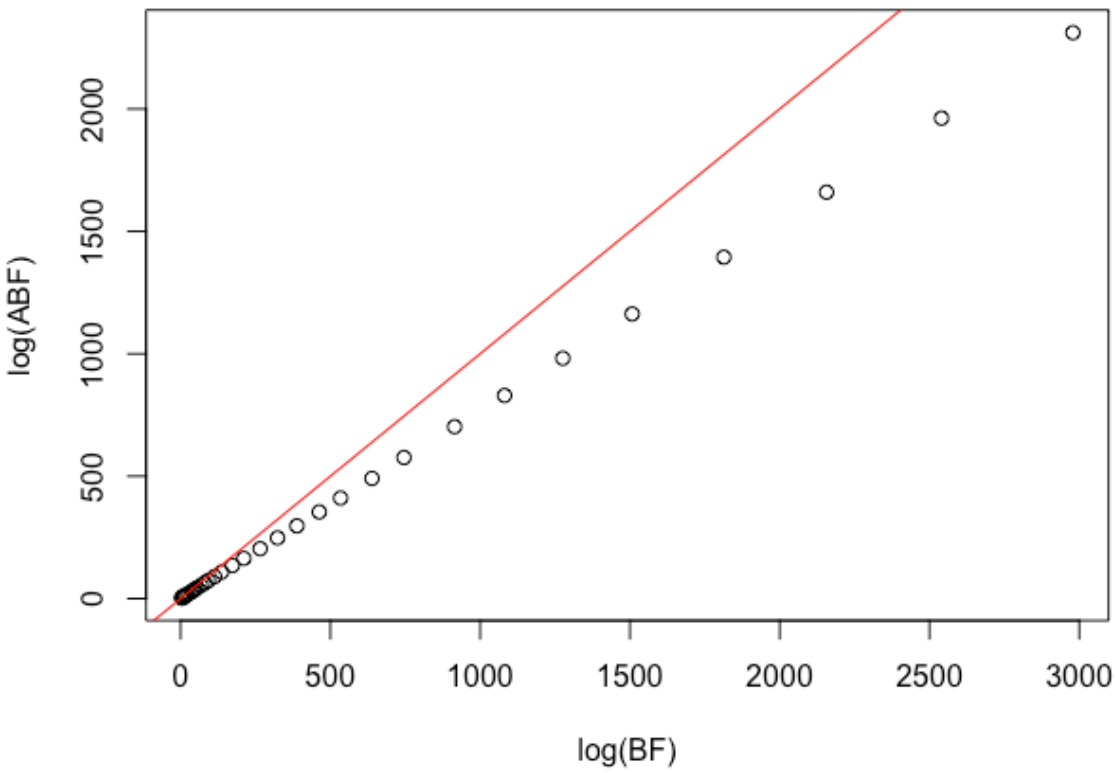

Figure 1: The ABF gets worse as the BF increases: we simulate data under a logistic regression model with the intercept = 0 and effect size 1. We increase the number of observations from $n = 2^6$ up to $n = 2^{15}$. We compute BFs assuming a mean zero normal prior on the effect size with variance 1. As the sample size increases, evidence against the null hypothesis becomes stronger. We see that the approximation error incurred by the ABF grows, being off by thousands of log-likelihood units at the largest sample size.

2. The Laplace approximation of the likelihood ratio is exact at $\beta = \hat{\beta}_{MLE}$. This is in contrast to the likelihood ratio approximation implied by ABF, which is exact at $\beta = 0$. The Laplace approximation of the likelihood ratio will result in a better approximation of the BF, because the approximation is more accurate in the regions that contribute most the the likelihood ratio.
3. Both approximations use the same approximaiton of the numerator, but different approximation of the denominator. So the Laplace approximation of the likelihood ratio can be though of as a rescaling of the likelihood ratio approximation used in the ABF (or translation, on the log scale) such that the approximate and exact likelihood ratios agree at at the MLE rather than at 0.

Example: variable selection

Here is an example of how using the ABF in the SER is problematic. But using the Laplace approximation works fine!

Discussion

1. The relative error ABF increases as the BF increases to (extremely) large values. This is because asymptotic approximation is not good in the tails ($\beta = 0$ is in the tail if the BF is large)
2. We could have avoided all of this fuss in the SER example if rather than using the asymptotic approximation of the BF (the ABF), we just used the asymptotic approximation of the marginal likelihood? If we were to approximate the BF between two alternative models we would just approximate the marginal likelihood of each and take their raito. Using the ratio of ABFs. introduces the added complication of differences in the approximation of the null model– which is not relevant to the comparison of the two alternative models.
3. The appeal of the ABF is that it only requires the effect size estimate and it's standard error. The proposed approximation also requires knowledge of the likelihood under the null model, or equivalently, the likelihood ratio between the MLE and the null model. This information can dramatically improve the approximation of the BF when the BF is large. This should not be surprising, however, if we think of our approximation of the BF as a Laplace approximation to the likelihood ratio. In this case, the Laplace approximation to the BF is centered on the exact likelihood ratio.

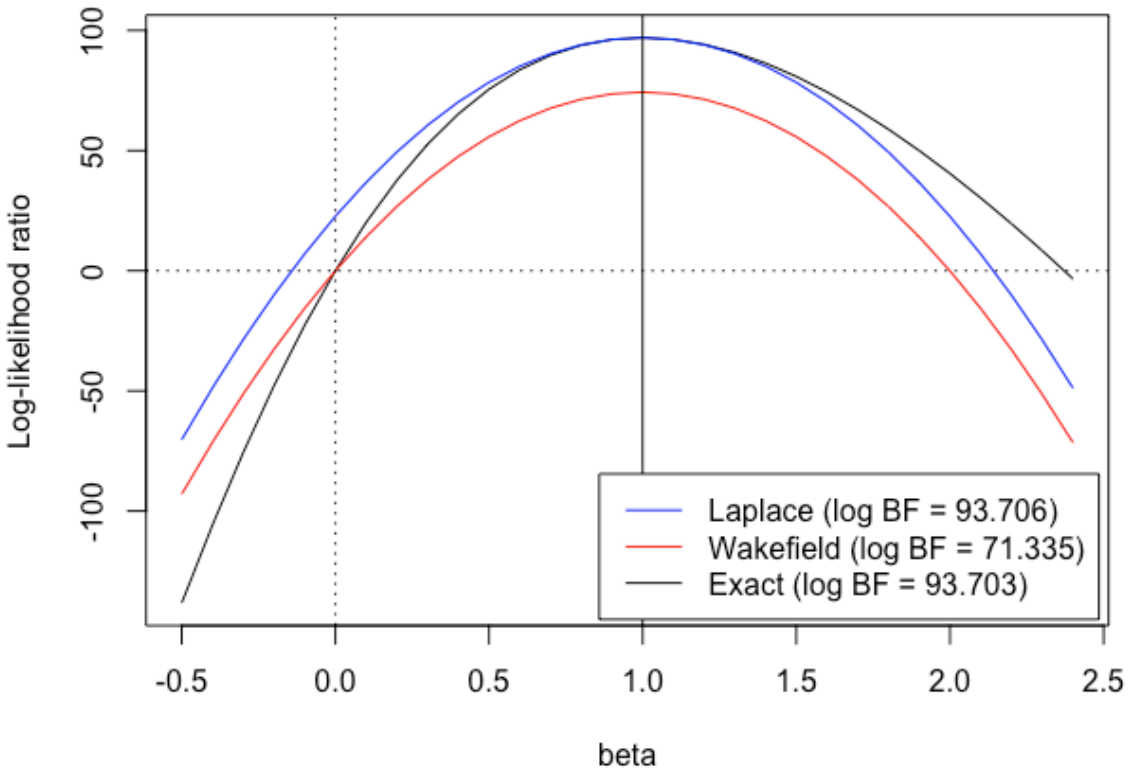

Figure 2: Likelihood ratio approximations. We simulate $n = 1000$ observations under a logistic regression model with $x_i \sim N(0, 1)$ and $\beta_0 = 0$ and $\beta = 1$. We plot the log likelihood ratio for varying β against the null model ($= 0$). We also plot the likelihood ratio approximations used in Wakefield's ABF and the proposed Laplace approximation. Wakefield and Laplace use the same approximation of the numerator of the BF but differ in the choice of denominator. Wakefield chooses to make the approximate likelihood ratio exact at $\beta = 0$ while Laplace makes the approximate likelihood ratio exact at $\beta = \hat{\beta}_{MLE}$. The latter gives better approximation the the BF, since the approximation is more accurate in areas that contribute most to the integral.

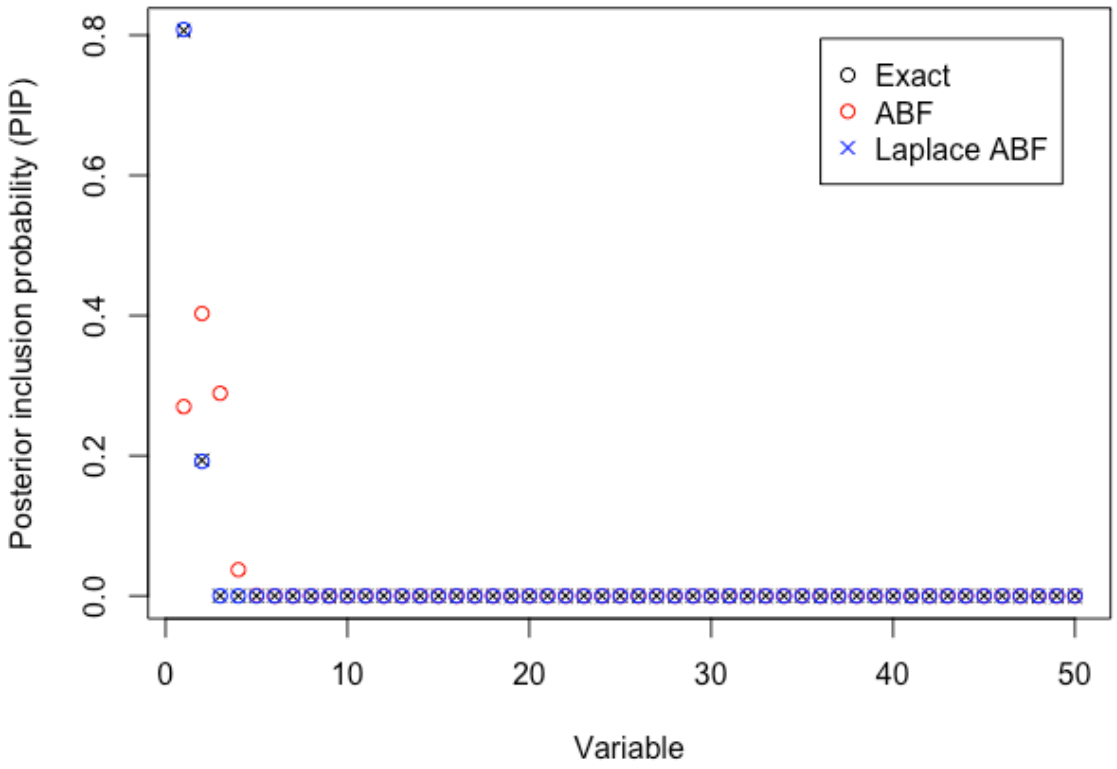

Figure 3: This is an example of the downstream consequences of using the ABF when having accurate/comparable BFs is important. Here we simulate $p = 50$ variables that are mean-zero normal but correlated. Only one variable (the first) has non-zero effect. We simulate with an intercept $\beta_0 = -2$, and an effect size $\beta = 1$ for the fist variable. Since the ABFs result in a different ordering of variables than the exact BFs, the SER will assign different posterior inclusiong probabilities to variables compared to the exact BFs. This is an artifact of each ABF having a different approximation of the denominator. If we use the Laplace approximate BF we get excellent agreement with the exact BFs, and therefor excellent agreement between the PIPs.