Bayes Factors for Informative Hypotheses

Jasper Ginn

April 8, 2019

Bayes Factors & Marginal Likelihoods

This part is based on [5]. Given two models, M_1 and M_2, and some data D we want to know the following probabilities:

(1) $p(M_1|D)$

(2) $p(M_2|D)$

These represent the probabilities of our model given the data. Note that (1) and (2) are posterior probabilities, such that (in the case of M_1), we have:

$$p(M_1|D) = \frac{p(D|M_1)p(M_1)}{p(data)} = \frac{\text{Marginal likelihood} \cdot \text{Model prior}}{\text{Maginal likelihood over } M_1 \text{ and } M_2} = \frac{p_1 p_2}{p_3} \tag{1}$$

1. **The marginal likelihood**

Recall that the posterior for some posterior θ is given as:

$$p(\theta|\text{data}) = \frac{p(\text{data}|\theta)p(\theta)}{p(\text{data})}$$

If we consider model 1, this is equivalent to:

$$p(\theta|\text{data}, M_1) = \frac{p(\text{data}|\theta, M_1)p(\theta|M_1)}{p(\text{data}, M_1)} \tag{2}$$

Notice that the denominator in equation (2) is the same as the marginal likelihood of equation (1). The problem is that this is essentially a difficult integral, especially if we are working in higher dimensions and θ contains multiple parameters:

1

$$p(\text{data}, M_1) = \int p(\text{data}|\theta, M_1)p(\theta|M_1)d\theta$$

2. **Marginal likelihood of M_1 and M_2**

This is a marginal density where we have integrated out the model choice, such that:

$$p(D) = p(D|M_1)p(M_1) + P(data|M_2) + p(M_2)$$

It is the same for each model and it contains information from **both** models.

3. **The model prior**

The prior on the model is often set at $p(M_1) = P(M_2) = \frac{1}{2}$ because, *a priori*, each model is equally likely.

Instead of calculating $p(M_1|D)$ and $p(M_2|D)$ separately (which is difficult), we can compute the *ratio* of one to the other, such that:

$$\frac{p(M_1|D)}{p(M_2|D)} = \frac{\frac{p(D|M_1)}{p(D)}}{\frac{p(D|M_2)}{p(D)}} \cdot \frac{P(M_1)}{P(M_2)} = \frac{p(D|M_1)}{p(D|M_2)} \cdot \frac{P(M_1)}{P(M_2)} = \text{BF}_{1,2}\frac{P(M_1)}{P(M_2)} \tag{3}$$

Hence, the Bayes' Factor (BF) gives us the odds of how our relative beliefs should change after seeing the data ("evidence"). The value of the BF depends on:

1. The strength of our prior beliefs
2. The strength of the posterior evidence w.r.t. the priors (our beliefs)
3. The choice of our prior
4. The complexity of our hypotheses/models

Bayes' Factors and Hypotheses

This part is based on [1], [2], [3], [4] and [6]. Consider two hypotheses:

1. $p(H_1|D)$
2. $p(H_u|D)$

Where H_1 is some statement about e.g. regression coefficients (for example $H_1 : \beta_1 > 0$) and H_u is the 'unconstrained' hypothesis ($H_u : \beta_1$), meaning 'there is some beta 1'. Then the BF 'weights' the prior belief (the hypotheses) against the evidence found in the data, interpred as the relative evidence in the data for one hypothesis relative to another exactly to the degree that the hypothesis predicts the observed data better than the other [6].

In the context of BF used by [1], [2], [3], [4], the Bayes Factor for some hypothesis H_1 against the unconstrained hypothesis H_u can be computed as:

$$\text{BF}_{1u} = \frac{\text{fit}_1}{\text{complexity}_1} = \frac{f_1}{c_1} \tag{4}$$

The BF for some hypothesis H_1 against its complement 'not H_1' can be computed as:

$$\text{BF}_{1c} = \frac{\frac{\text{fit}_1}{\text{complexity}_1}}{\frac{1-\text{fit}_1}{1-\text{complexity}_1}} = \frac{\frac{f_1}{c_1}}{\frac{1-f_1}{1-c_1}} \tag{5}$$

Where:

1. The **fit** of a hypothesis is the proportion of the posterior distribution that is supported by the hypothesis.
2. The **complexity** of a hypothesis is the proportion of the prior distribution that is supported by the hypothesis.

Example in R

Assume that we have a linear regression equation with two coefficients, β_1 and β_2. We are fairly uncertain about our prior knowledge and hence we set $\beta_1, \beta_2 \sim N(0, 100)$

```
# Number of ficticious iterations for an MCMC sampler
k <- 20000


# Generate prior data
b1_prior <- rnorm(k, 0, 100)
b2_prior <- rnorm(k, 0, 100)
```

```r
# Outcome variable y
y <- rnorm(100, 3.2 + 5*rnorm(100, 8, 1.5) + 8*rnorm(100, -5, 2), sd=3.4)
```

We have the following three hypotheses:

1. $H_1 : \beta_1 > \beta_2$
2. $H_2 : \beta_1 \approx \beta_2 = |\beta_1 - \beta_2| < .1$
3. $H_u : \beta_1, \beta_2$

In order to compare the prior and posterior distributions, we need to standardize the coefficients. This, we can do using

$$\beta_1^* = \beta_1 \cdot \sigma_{\beta_1} / \sigma_y \tag{6}$$

Then for the prior distributions, we get the following:

```r
library(ggplot2)
library(gridExtra)
library(latex2exp)


# Draw prior data
priorD <- data.frame(
  "b1pr" = b1_prior * sd(b1_prior) / sd(y),
  "b2pr"= b2_prior * sd(b1_prior) / sd(y)
)


# Plot
p1 <- ggplot(priorD, aes(x=b1pr, y=b2pr)) +
  geom_density2d() +
  geom_abline(intercept = 0, slope=1, linetype="dashed") +
  blm::theme_blm() +
  theme(axis.text = element_blank(),
        axis.ticks = element_blank()) +
```

```r
  scale_x_continuous(name = TeX("$\\beta_{1, prior}$")) +

  scale_y_continuous(name = TeX("$\\beta_{2, prior}$"))


# Plot

p2 <- ggplot(priorD, aes(x=b1pr, y=b2pr)) +

  geom_density2d() +

  geom_abline(intercept = -250, slope=1, linetype="dashed") +

  geom_abline(intercept = +250, slope=1, linetype="dashed") +

  blm::theme_blm() +

  theme(axis.text = element_blank(),

        axis.ticks = element_blank())+

  scale_x_continuous(name = TeX("$\\beta_{1, prior}$")) +

  scale_y_continuous(name = TeX("$\\beta_{2, prior}$"))


grid.arrange(p1, p2, ncol=2)
```
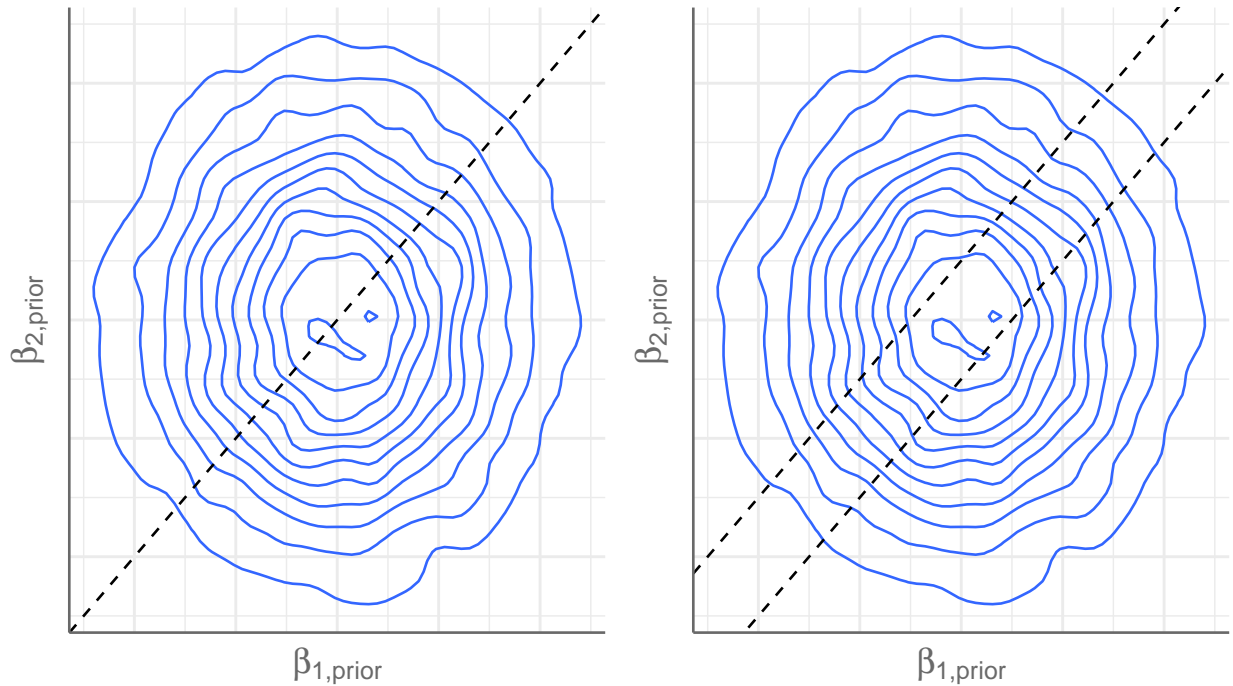

Figure 1: In the left plot, evething under the dashed line corresponds to H1. In the right plot, the area between the dashed lines correspond to H2.

```r
# Complexity
c_1 <- mean(priorD$b1pr > priorD$b2pr)

c_2 <- mean(abs(priorD$b1pr - priorD$b2pr) < 250)
```

From the image above, we should notice that the *complexity* of the hypotheses, which in the case of the left plot is everything under the dashed line and area in between the shaded lines for the right plot, vastly differs. In other words, H_1 encompasses a larger area and is therefore more *complex* because we allow for more possibilities. The complexity for hypothesis 1 is $c_1 = p(\beta_1 > \beta_2) \approx .49$ and for hypothesis 2 it is $c_2 = p(|\beta_1 - \beta_2| < .1) \approx .25$

Now assume that we have a posterior distribution for the parameters. $\beta_{1,\text{posterior}} \sim N(5, 2)$ and $\beta_{2,\text{posterior}} \sim N(8, 4)$. Then:

```r
# Simulate posterior distributions
b1_posterior <- rnorm(k, 5, 2)
b2_posterior <- rnorm(k, 8, 4)


# To df
postD <- data.frame(
  "b1po" = b1_posterior * sd(b1_posterior) * sd(y),
  "b2po" = b2_posterior * sd(b2_posterior) * sd(y)
)


# Plot
p1 <- ggplot(postD, aes(x=b1po, y=b2po)) +
  geom_density2d() +
  geom_abline(intercept = 0, slope=1, linetype="dashed") +
  blm::theme_blm() +
  theme(axis.text = element_blank(),
        axis.ticks = element_blank()) +
  scale_x_continuous(name = TeX("$\\beta_{1, posterior}$")) +
  scale_y_continuous(name = TeX("$\\beta_{2, posterior}$"))


# Plot
```
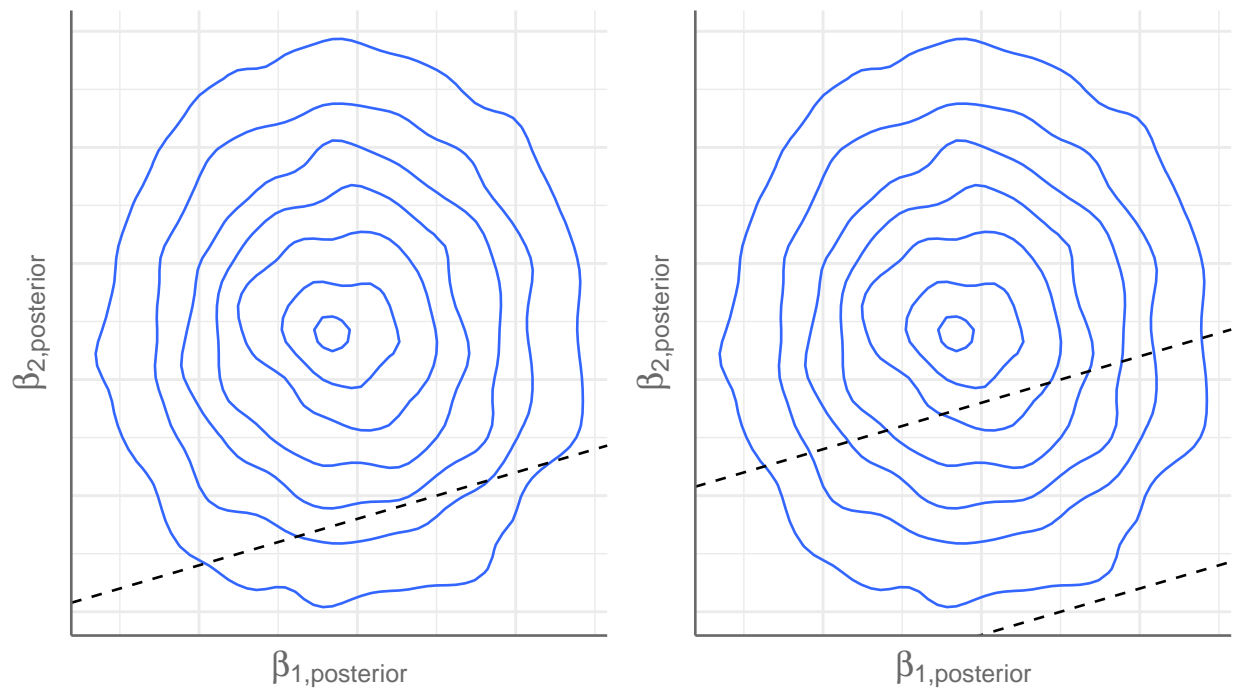
```r
p2 <- ggplot(postD, aes(x=b1po, y=b2po)) +
  geom_density2d() +
  geom_abline(intercept = -250, slope=1, linetype="dashed") +
  geom_abline(intercept = +250, slope=1, linetype="dashed") +
  blm::theme_blm() +
  theme(axis.text = element_blank(),
        axis.ticks = element_blank())+
  scale_x_continuous(name = TeX("$\\beta_{1, posterior}$")) +
  scale_y_continuous(name = TeX("$\\beta_{2, posterior}$"))


grid.arrange(p1, p2, ncol=2)
```



```r
# Fit
f_1 <- mean(postD$b1po > postD$b2po)
f_2 <- mean(abs(postD$b1po - postD$b2po) < 250)


# BF against each other
```

```
H1vH2 <- (f_2 / c_2) / (f_1 / c_1)
```

From the above plot, we see that the area corresponding to the first hypothesis (left) is smaller than the area corresponding to the first hypothesis of the prior. When we calculate the BF against the unconstricted hypothesis, we get $f_1/c_1 \approx .19$. For the second hypothesis, we get $f_2/c_2 \approx 1.15$. This means that, after seeing the data, we now favor H_1 over H_u by a factor of .19, whereas we favor H_2 over H_u by 1.15. We can also compare H_1 and H_2, which yields $\frac{f_1/c_1}{f_2/c_2} \approx 6.18$, meaning that we favor H_2 over H_1 by a factor of 6.18.

References

[1] Hoijtink, H., Mulder, J., Van Lissa, C. J., & Gu, X. (2019). A tutorial on testing hypotheses using the Bayes factor.

[2] Hoijtink, H., Gu, X., & Mulder, J. (2018). Bayesian evaluation of informative hypotheses for multiple populations. British Journal of Mathematical and Statistical Psychology.

[3] Gu, X., Mulder, J., & Hoijtink, H. (2018). Approximated adjusted fractional Bayes factors: A general method for testing informative hypotheses. British Journal of Mathematical and Statistical Psychology, 71(2), 229-261.

[4] Gu, X., Hoijtink, H., & Mulder, J. (2016). Error probabilities in default Bayesian hypothesis testing. Journal of Mathematical Psychology, 72, 130-143.

[5] Lambert, B. [Ben Lambert]. (2018, May 17). *Introducing Bayes Factors and Marginal Likelihoods*. Retrieved from https://www.youtube.com/watch?v=T-kMpA4z-7k

[6] Morey, R. (2014, February 9). *What is a Bayes Factor?*. [Blog Post]. Retrieved from https://bayesfactor. blogspot.com/2014/02/the-bayesfactor-package-this-blog-is.html