Bayes: Homework 10
Leslie Gains-Germain

1. Here are some comments and questions. I'm not expecting you to answer these. Many of these were just ways for me to work and think through what was going on.

   The effective number of parameters is the average deviance (found for all posterior draws of θ) minus the deviance of the posterior mean of θ. I'm having trouble thinking about what this quantity means, and if it's just meant to give us an idea of the posterior variance of the deviance, why is it called the effective number of parameters? Also called the bias correction for overfitting?

   Steps for finding DIC:

   (a) Find deviance = -2*log likelihood of data given each posterior draw for θ.

   (b) Find the average of these deviances (deviance.avg)

   (c) Calculate deviance.postmean = -2log likelihood of data given the posterior mean of θ.

   (d) Calculate pD = deviance.avg - deviance.postmean

   (e) Find DIC = deviance.avg + pD

   Why is the deviance more spread out for the no pooling model? Maybe because the θ's themselves are more spread out?

   Complete pooling model has the lowest DIC - which Gelman discusses on pages 179-180. I think this relates to something we talked about in class one day - if we used ANOVA to assess variability among groups, we would find no evidence for a difference among the group effects.

When would you want to use the deviance of the yreps vs the posterior predicted MSE?

How often is AIC actually used in a Bayesian setting? It seems like Gelman introduces it, but says it's not really applicable if we go beyond linear models with flat priors. Link and Barker talk about it being used by fields of wildlife biology and ecology fields, but I'm not sure if they mean in a Bayesian setting or not. It definitely seems easier to calculate!

I remember being taught in 502 that the estimate for θ is the posterior mean under squared error loss, and the estimate for θ under absolute error loss is the posterior median. Why?

Why do AIC, DIC, and WAIC multiply by -2? It seems like they could just be defined as the log predictive densities minus the bias correction. It seems like the -2 is a connection to the asymptotic results for likelihood ratio tests, but when comparing AICs, DICS or WAICS, is it really necessary? I guess Gelman talks a lot about putting these measurees "on the deviance scale". Maybe this is a silly question, I guess this is just how deviance is defined, and deviance residuals are defined this way too.

AIC, DIC, and WAIC focused on estimation of predictive fit, BIC (and I think Bayes Factors too) focused on estimating relative posterior probabiilities in a setting of discrete model comparison

Model averaging incorporates model weights in parameter estimation to account for model uncertainty.

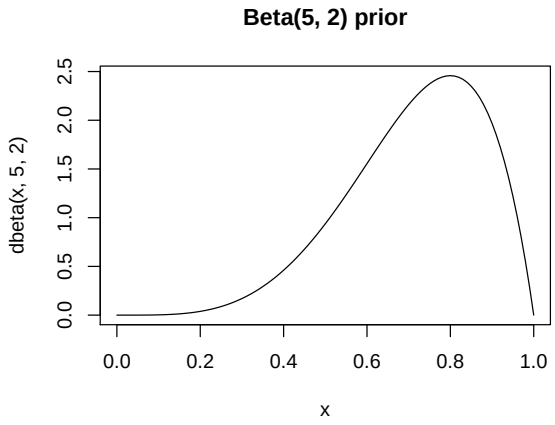
Is model averaging ever used in a frequentist context?

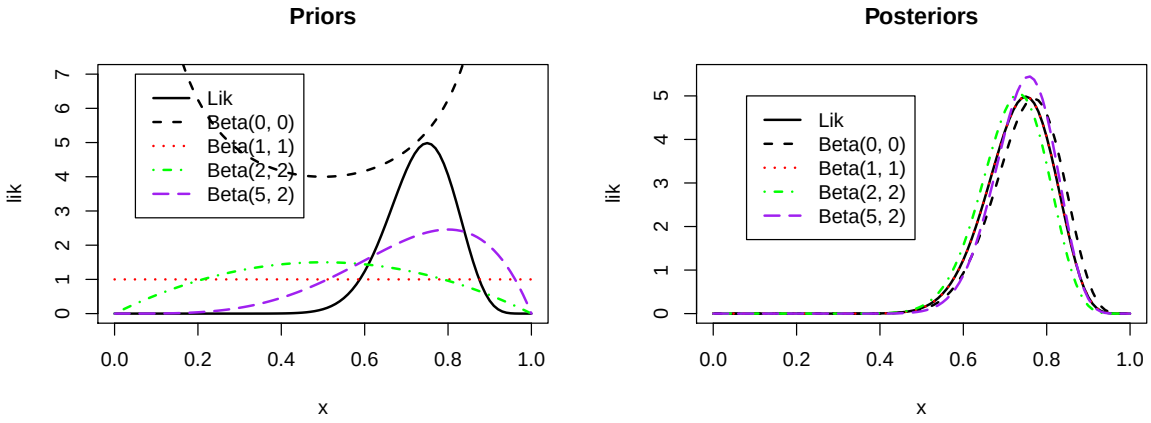
Is model averaging a way to avoid choosing a model?

How important is model averaging? What if you use Bayesian inference but never use multi-model inference? Is that bad?

In some cases, it seems like model choice isn't relevant. For example, if you know the data generating process followed a binomial distribution, and you put all your prior knowledge into the model, what would you gain from comparing to other models? (maybe Gelman's point in `Gelman_Rubin_discussRaftery` paper?)

2. (a) I will use the analytical results derived previously in the course. For a Binomial likelihood and a $Beta(\alpha, \beta)$ prior, the posterior distribution for π (the true probability that Liz's cupcakes taste better than Megan's) is $Beta(y + \alpha, n - y + \beta)$. So, the $Beta(0, 0)$ prior results in a $Beta(21, 7)$ posterior. The $Beta(1, 1)$ prior results in a $Beta(22, 8)$ posterior. The $Beta(2, 2)$ prior results in a $Beta(23, 9)$ posterior. I chose a $Beta(5, 2)$ for the informative prior because it gives higher density to probabilities between 0.6 and 0.9. I've never had Megan's cupcakes so I don't want to rule out the possibility that her's are better, but I think Liz's cupcakes would be hard to beat. Because of different tastes, however, I think there will always be some people who prefer Megan's, and this is why I have chosen a prior with lower density near 1 (see plot below). With a $Beta(5, 2)$ prior, the posterior is $Beta(26, 9)$.





When the informative $Beta(5, 2)$ prior is used, the posterior chance that the probability a Liz-made cupcake is perceived to be tastier than a Megan-made cupcake is estimated to be 0.758 (posterior mode), with a 95% posterior interval from 0.588 to 0.871. The posterior chance that the probability a Liz-made cupcake is perceived to be better is greater than 0.5 is 0.998.

(b) I show my work for finding the Bayes Factor for the $Beta(0, 0)$ prior below. I used similar

calculations for the other priors.

$$\begin{aligned}\frac{p(M_1|y)/p(M_2|y)}{p(M_1)/p(M_2)} &= \frac{\frac{p(y|M1)p(M_1)/p(y)}{p(y|M_2)p(M_2)/p(y)}}{p(M_1)/p(M_2)} = \frac{p(y|M_1)}{p(y|M_2)} = \frac{\int_{\theta|M_1} p(y|M_1,\theta)p(\theta|M_1)}{\int_{\theta|M_2} p(y|M_2,\theta)p(\theta|M_2)} \\ &= \frac{\int_{0.7}^{1} \theta^y(1-\theta)^{28-y}\theta^{-1}(1-\theta)^{-1}d\theta}{\int_{0}^{0.7} \theta^y(1-\theta)^{28-y}\theta^{-1}(1-\theta)^{-1}d\theta} \\ &= \frac{\int_{0.7}^{1} \theta^{21-1}(1-\theta)^{7-1}}{\int_{0}^{0.7} \theta^{21-1}(1-\theta)^{7-1}} = 2.9\end{aligned}$$

	priors	BayesFactors
1	Beta(0, 0)	2.90
2	Beta(1, 1)	2.11
3	Beta(2, 2)	1.59
4	Beta(5, 2)	2.74

With a $Beta(0, 0)$ prior, the bayes factor is 2.9. Assuming neither model was favored a priori, the posterior odds that the probability a Liz-made cupcake is perceived to be tastier is greater than 0.7 is 2.9. In the remaining summary sentences, I assume neither model was favored a priori. With a $Beta(1, 1)$ prior, the posterior odds that the probability a Liz-made cupcake is perceived to be tastier is greater than 0.7 is 2.1. With a $Beta(2, 2)$ prior, the posterior odds that the probability a Liz-made cupcake is perceived to be tastier is greater than 0.7is 1.59. With a $Beta(5, 2)$ prior, the posterior odds that the probability a Liz-made cupcake is perceived to be tastier is greater than 0.7 is 2.74.

3. Priors matter for Bayes Factors because the priors on the parameters are directly involved in the formula for a Bayes Factor ($p(\theta|M_1)$ and $p(\theta|M_2)$ in the formulas below).

$$BF = \frac{p(M_1|y)/p(M_2|y)}{p(M_1)/p(M_2)} = \frac{\frac{p(y|M1)p(M_1)/p(y)}{p(y|M_2)p(M_2)/p(y)}}{p(M_1)/p(M_2)} = \frac{p(y|M_1)}{p(y|M_2)} = \frac{\int_{\theta|M_1} p(y|M_1,\theta)p(\theta|M_1)}{\int_{\theta|M_2} p(y|M_2,\theta)p(\theta|M_2)}$$

The prior used for θ will directly affect the value of the Bayes Factor. Bayes Factors are not always helpful when comparing models, especially when vague priors are used. Consider an

example where model 1 is $y \sim N(0,1)$, and model 2 is $y \sim N(\mu,1)$ with $\mu \sim N(\mu_0, \tau^2)$.

$$BF = \frac{p(y|M_1)}{p(y|M_2)} = \frac{\frac{1}{\sqrt{2\pi}} e^{-y^2/2}}{\int_\mu \frac{1}{\sqrt{2\pi}} e^{-(y-\mu)^2/2} \frac{1}{\sqrt{2\pi\tau^2}} e^{-(\mu-\mu_0)^2/2\tau^2} d\mu}$$

It is clear from the above equation that as τ goes to infinity, the denominator goes to 0, and the entire equation goes to ∞. So, as the prior variance increases, more weight is given to the more informative $y \sim N(0,1)$ model. So, if we chose τ to be large and used the Bayes factor blindly, we would choose the $N(0,1)$ model. If we use common sense in this situation, however, we would realize that model 2 is the better choice. If we chose a large value of τ for model 2 originally, then we are uncertain about the true value of μ. This means that model 1 is not a very good choice, because it reflects no uncertainty in μ. Model 2 is a much better choice because it does incorporate uncertainty in μ. This example shows that Bayes Factors are not always helpful when selecting models, and trusting Bayes Factors without using common sense can be dangerous, especially when vague priors are used.

4. A likelihood ratio is a special case of a Bayes Factor. Suppose we are testing two models, $M_1 : y \sim N(0,1)$ vs $M_2 : y \sim N(1,1)$. In this situation, the Bayes factor is the same as the likelihood ratio if neither model is favored a priori. It is simply a ratio of likelihood functions at $\mu = 0$ and $\mu = 1$.

$$BF = LR = \frac{p(y|M_1)}{p(y|M_2)} = \frac{\frac{1}{\sqrt{2\pi}} e^{-y^2/2}}{\frac{1}{\sqrt{2\pi}} e^{-(y-1)^2/2}}$$

I think of the Bayes Factor as being more flexible because it allows us to incorporate uncertainty in model parameters into our comparison by integrating over all values of the parameter. For an example of this, refer to the denominator of the Bayes Factor in the previous problem. Bayes Factors can also be used to compare two models of different forms, such as the poisson and geometric models in problem 5. In a likelihood ratio test, the models being compared have to have the same form (both poisson, both binomial, etc).

5. (a) i. Assuming neither model is favored a priori, the prior odds for the geometric model are $0.5/0.5 = 1$.

ii. The posterior odds that the geometric model is the data generating model are 0.200. (So the odds the poisson model is the data generating model are 5.01). My work is shown below.

$$\begin{aligned}\frac{p(M_1|y)}{p(M_2|y)} &= \frac{p(y|M_1)p(M_1)/p(y)}{p(y|M_2)p(M_2)/p(y)} = \frac{p(y|M_1)}{p(y|M_2)} \\ &= \frac{0.15^5(1-0.15)^{23}}{\frac{5^{23}e^{-25}}{1!3!5!7!7!}}\end{aligned}$$

iii. Since the prior odds are equal, the Bayes Factor is equal to the posterior odds, 0.200.

(b) i. The prior odds for the geometric model are still 1 because neither model is favored a priori.

ii. The posterior odds for the geometric model are 0.594. My work is shown below. I recognized a beta kernel in the numerator and gamma kernel in the denominator. I show my code below for the integration.

$$\begin{aligned}\frac{p(M_1|y)}{p(M_2|y)} &= \frac{p(y|M1)p(M_1)/p(y)}{p(y|M_2)p(M_2)/p(y)} = \frac{p(y|M_1)}{p(y|M_2)} \\ &= \frac{\int_{\pi|M_1} p(y|M_1,\pi)p(\pi|M_1)}{\int_{\lambda|M_2} p(y|M_2,\lambda)p(\lambda|M_2)} \\ &= \frac{\int_{1/31}^{1} \pi^5(1-\pi)^{23}\frac{1}{30\pi^2}}{\int_0^{30}\frac{\lambda^{23}e^{-5\lambda}}{1!3!5!7!7!}\frac{1}{30}} \\ &= \frac{\int_{1/31}^{1} \pi^3(1-\pi)^{23}}{\int_0^{30}\frac{\lambda^{23}e^{-5\lambda}}{1!3!5!7!7!}} = 0.594\end{aligned}$$

```
num <- (pbeta(1,4,24) - pbeta(1/31, 4, 24))*gamma(4)*gamma(24)/gamma(28)

facs <- (factorial(3)*factorial(5)*factorial(7)*factorial(7))

gam <- (pgamma(30, shape = 24, scale = 1/5) - pgamma(0, shape = 24, scale = 1/5))*
  gamma(24)*(1/5)^24
denom <- gam/facs
```

```
postodds2 <- num/denom
```

iii. Again, because we don't favor one model a priori, the bayes factor is the same as the posterior odds for the geometric model, 0.594.