BE/Bi 103: Data Analysis in the Biological Sciences

Justin Bois

Caltech

Fall, 2018

This document was prepared at Caltech with financial support from the Donna and Benjamin M. Rosen Bioengineering Center.

7 Hierarchical models

In this lecture, we will investigate **hierarchical models**, in which some model parameters are dependent on others in specific ways. This is best learned by example.

In homework problem 6.1, we studied reversals under exposure to blue light in *C. elegans* with Channelrhodopsin in two different neurons. Let's consider one of the strains which contains a Channelrhodopsin in the ASH sensory neuron. We considered data done in three different years by the students of Bi 1x. In 2015, we found that 9 out of 35 worms reversed under exposure to blue light. In 2016, 12 out of 35 reversed. In 2017, 18 out of 54 reversed.

7.1 Analytical expression for the posterior

This is one of the few examples where we can write an analytical expression for the posterior. We will do so because it will make discussion about hierarchical models simpler.

Consider for a moment only a single experiment, we can use this measurement to estimate the probability θ of reversal. In the homework, you modeled the number of reversals with a Binomial distribution and the probability of reversal θ with a Beta distribution.

$$\theta \sim \text{Beta}(\alpha, \beta), \tag{7.1}$$

$$n \sim \text{Binom}(N, \theta), \tag{7.2}$$

where n is the number of reversals and θ is the probability of reversal upon exposure to blue light. This parameter θ is what we wish to estimate.

We can write out this model in full detail using Bayes's theorem.[11]

$$g(\theta \mid n, N) = \frac{f(n \mid N, \theta)\, g(\theta)}{f(n \mid N)}, \tag{7.3}$$

where

$$f(n \mid N, \theta) = \frac{N!}{(N-n)!n!}\, \theta^n (1-\theta)^{N-n}, \tag{7.4}$$

and

$$g(\theta) = \frac{1}{B(\alpha, \beta)}\, \theta^{\alpha-1}(1-\theta)^{\beta-1}, \tag{7.5}$$

[11] Note that I wrote $g(\theta)$ instead of $g(\theta \mid N)$ because they are equal; N has no bearing on θ.

where $B(\alpha,\beta) = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha+\beta)$ is the Beta function. Putting this all together enables us to write an expression for the posterior.

$$g(\theta \mid n, N) = \frac{1}{f(n \mid N) B(\alpha,\beta)} \frac{N!}{(N-n)!n!} \left[\theta^{n+\alpha-1}(1-\theta)^{N-n+\beta-1}\right]. \tag{7.6}$$

Looking at this expression, the bracketed expression is the only bit that depends on θ. This is exactly the θ-dependence of a Beta distribution with parameters $n+\alpha$ and $N-n+\beta$. Because the posterior must be normalized, the posterior therefore must be a Beta distribution and

$$\frac{1}{f(n \mid N)\, B(\alpha,\beta)} \frac{N!}{(N-n)!n!} = \frac{1}{B(n+\alpha, N-n+\beta)}. \tag{7.7}$$

We have just normalized the posterior without doing any nasty integrals! So, the posterior is

$$g(\theta \mid n, N) = \frac{\theta^{n+\alpha-1}(1-\theta)^{N-n+\beta-1}}{B(n+\alpha, N-n+\beta)}, \tag{7.8}$$

or,

$$\theta \mid n, N \sim \text{Beta}(n+\alpha, N-n+\beta). \tag{7.9}$$

It is clear that the data have updated the parameters of the Beta prior.

For a given likelihood, a prior that is the same distribution as the posterior (obviously with different parameters) is said to be **conjugate** to the likelihood. So, we can see that conjugacy is useful. For a given likelihood, if we know its conjugate prior, we can just immediately write down the posterior in a clear form. The Wikipedia page on conjugate priors has a useful table of likelihood-conjugate pairs.

Note though that a closed form conjugate does not always exist for a given likelihood, especially for complicated models, and when they do exist, they may be very difficult to find. This does limit their utility. Further, there is no reason why a posterior and prior should have the same functional form; all analysis is completely valid without conjugacy.

7.2 Ways to model repeated experiments

We have the posterior for a single experiment. But, we did the experiment in 2015, getting $n/N = 9/35$, and again in 2016, getting $n/N = 12/35$, and in 2017 with $n/N = 18/54$. Actually, we could imagine doing the experiment over and over again, say k times, each time getting a value of n and N. Conditions may change from experiment to experiment. For example, we may have different lighting setups, slight differences in the strain of worms we're using, etc. We are left with some choices on how to model the data.

7.2.1 Pooled data: identical parameters

We could pool all of the data together. In other words, let's say we measure n_1 out of N_1 reversals in the first set of experiments, n_2 out of N_2 reversals in the second set, etc., up to k total experiments. We could pool all of the data together to get

$$n = \sum_{i=1}^{k} n_i \text{ out of } \sum_{i=1}^{k} N_i \text{ reversals}. \tag{7.10}$$

We then compute our posterior as in equation (7.8). Here, the modeling assumption is that the result in each experiment are governed by *identical parameters*. That is to say that we assume $\theta_1 = \theta_2 = \cdots = \theta_k = \theta$.

This is similar to what we did in homework problem 2.3, in which showed how a single hypothesis (or parameter value) is informed by more data. And this is the modeling approach we took in homework problem 6.1.

7.2.2 Independent parameters

As an alternative model, we could instead say that the parameters in each experiment are totally independent of each other. In this case, we assume that θ_1, θ_2, ..., θ_k are all independent of each other. The likelihoods and priors all multiply and the posterior probability is

$$\theta_i \sim \text{Beta}(n + \alpha, N - n + \beta) \text{ for all } i. \tag{7.11}$$

When we make this assumption, we often report a value of θ that is given by the mean of the θ_i's with some error bar.

7.3 Best of both worlds: a hierarchical model

Each of these extremes have their advantages. We are often trying to estimate a parameter that is more universal than our experiments, e.g., something that describes worms with Channelrhodopsin in the ASH neuron generally. So, pooling the experiments makes sense. On the other hand, we have reason to assume that there is going to be a different value of θ in different experiments, as biological systems are highly variable, not to mention measurement variations. So, how can we capture both of these effects?

We can consider a model in which there is a "master" reversal probability, which we will call ϕ, and the values of θ_i may vary from this ϕ according to some probability distribution, $g(\theta_i \mid \phi)$. So now, we have parameters $\theta_1, \theta_2, \ldots, \theta_k$ and ϕ.

So, the posterior can be written using Bayes's theorem, defining $\theta = (\theta_1, \theta_2, \ldots)$, $N = (N_1, N_2, \ldots)$, and $n = (n_1, n_2, \ldots)$,

$$g(\phi, \theta \mid n, N) = \frac{f(n, N \mid \phi, \theta)\, g(\phi, \theta)}{f(n, N)}. \tag{7.12}$$

Note, though, that the observed values of n do not depend directly on ϕ, only on θ. In other words, the observations are only *indirectly* dependent on ϕ. So, we can write $f(n, N \mid \phi, \theta) = f(n, N \mid \theta)$. Thus, we have

$$g(\phi, \theta \mid n, N) = \frac{f(n, N \mid \theta)\, g(\phi, \theta)}{f(n, N)}. \tag{7.13}$$

Next, we can rewrite the prior using the definition of conditional probability.

$$g(\phi, \theta) = g(\theta \mid \phi)\, g(\phi). \tag{7.14}$$

Substituting this back into our expression for the posterior, we have

$$g(\phi, \theta \mid n, N) = \frac{f(n, N \mid \theta)\, g(\theta \mid \phi)\, g(\phi)}{f(n, N)}. \tag{7.15}$$

Now, if we read off the numerator of this equation, we see a chain of dependencies. The experimental results n depend on parameters θ. Parameters θ depend on *hyperparameter* ϕ. Hyperparameter ϕ then has some **hyperprior** distribution. Any model that can be written as a chain of dependencies like this is called a **hierarchical model**, and the parameters that do not *directly* influence the data are called **hyperparameters**.

So, the hierarchical model captures both the experiment-to-experiment variability, as well as the master regulator of outcomes. Note that the product $g(\theta \mid \phi)\, g(\phi)$ comprises the prior, as it is independent of the observed data.

7.4 Exchangeability

The conditional probability, $g(\theta \mid \phi)$, can take any reasonable form. In the case where we have no reason to believe that we can distinguish any one θ_i from another prior to the experiment, then the label "i" applied to the experiment may be exchanged with the label of any other experiment. I.e., $g(\theta_1, \theta_2, \ldots, \theta_k \mid \phi)$ is invariant to permutations of the indices. Parameters behaving this way are said to be **exchangeable**. A common (simple) exchangeable distribution is

$$g(\theta \mid \phi) = \prod_{i=1}^{k} g(\theta_i \mid \phi), \tag{7.16}$$

which means that each of the parameters is an independent sample out of a distribution $g(\theta_i \mid \phi)$, which we often take to be the same for all i. This is reasonable to do in the worm reversal example.

7.5 Choice of the conditional distribution

We need to specify our prior, which for this hierarchical model means that we have to specify the conditional distribution, $g(\theta_i \mid \phi)$, as well as $g(\phi)$. We could assume a Beta prior for ϕ; the one you used in your homework for the reversal probability would be a good choice. For the conditional distribution $g(\theta_i \mid \phi)$, we might assume it is Beta-distributed. This necessitates another parameter because the Beta distribution has two parameters.

The Beta distribution is typically written as

$$g(\theta \mid \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)}\, \theta^{\alpha-1}(1-\theta)^{\beta-1}, \tag{7.17}$$

where it is parametrized by positive constants α and β. The Beta distribution has mean and **concentration**, respectively, of

$$\phi = \frac{\alpha}{\alpha + \beta}, \tag{7.18}$$

$$\kappa = \alpha + \beta. \tag{7.19}$$

The concentration κ is a measure of how sharp the distribution is. The bigger κ is, the most sharply peaked the distribution is.

Since we would like to parametrize our Beta distribution with its mean ϕ, we could use κ as our other parameter. So, our expression for the posterior is

$$g(\theta, \phi, \kappa \mid n, N) = \frac{f(n, N \mid \theta)\, \left(\prod_{i=1}^{k} g(\theta_i \mid \phi, \kappa)\right)\, g(\phi, \kappa)}{f(n, N)}. \tag{7.20}$$

We are left to specify the hyperprior $g(\phi, \kappa)$. We will take ϕ to come from a Beta distribution and κ to come from an weakly informative Half-Normal. Note that to switch from a parametrization using ϕ and κ to one using α and β, we can use

$$\alpha = \phi\kappa \tag{7.21}$$

$$\beta = (1 - \phi)\kappa. \tag{7.22}$$

With all of this, we can now put together our model.

$$\phi \sim \text{Beta}(2, 2), \tag{7.23}$$

$$\kappa \sim \text{HalfNorm}(0, 10), \tag{7.24}$$

$$\alpha = \phi\kappa, \tag{7.25}$$

$$\beta = (1-\phi)\kappa, \tag{7.26}$$

$$\theta_i \sim \text{Beta}(\alpha, \beta)\ \ \forall i, \tag{7.27}$$

$$n_i \sim \text{Binom}(N_i, \theta_i)\ \ \forall i. \tag{7.28}$$

7.6 Implementation

The lore has it that the original motivation for Stan's development was to enable effective sampling of hierarchical models (naive samplers often fail spectacularly at hierarchical models). To see the worm reversal problem solved with a hierarchical model, see the implementation here, wherein Stan shreds.

7.7 Generalization

The worm reversal problem is easily generalized. You can imagine having more levels of the hierarchy. This is just more steps in the chain of dependencies that are factored in the prior. For general parameters θ and hyperparameters ϕ, we have, for data set y,

$$g(\theta, \phi \mid y) = \frac{f(y \mid \theta)\, g(\theta \mid \phi)\, g(\phi)}{f(y)} \tag{7.29}$$

for a two-level hierarchical model. For a three-level hierarchical model, we can consider hyperparameters ξ that depend on ϕ, giving

$$g(\theta, \phi, \xi \mid y) = \frac{f(y \mid \theta)\, g(\theta \mid \phi)\, g(\phi \mid \xi)\, g(\xi)}{f(y)}, \tag{7.30}$$

and so on for four, five, etc., level hierarchical models. As we have seen in the course, the work is all in coming up with the models for the likelihood $f(y \mid \theta)$, and prior, $g(\theta \mid \phi)\, g(\phi)$, in this case for a two-level hierarchical model. For coming up with the conditional portion of the prior, $g(\theta \mid \phi)$, we often assume a Gaussian distribution because this often describes experiment-to-experiment variability. (The Beta distribution we used in our example is approximately Gaussian and has the convenient feature that it is defined on the interval $[0, 1]$.) Bayes's theorem gives you the posterior, and it is then "just" a matter of computing it by sampling from it. In coming tutorials, we will use Stan to sample out of hierarchical models and discuss the difficulties involved with doing that.