BE/Bi 103: Data Analysis in the Biological Sciences

Justin Bois

Caltech

Fall, 2017

9 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 5.2, 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. The experiment was performed three times 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.

Considering for a moment only the 2015 experiment, we can use this measurement to estimate the probability p of reversal. We modeled the likelihood of reversal with a Binomial likelihood. Taking a uniform prior on p, we derived that the posterior probability of reversal given r our of n trials showed reversals was

$$g(p \mid r, n) = \begin{cases} \dfrac{(n+1)!}{(n-r)!r!}\, p^r (1-p)^{n-r} & 0 \le p \le 1 \\ 0 & \text{otherwise.} \end{cases} \tag{9.1}$$

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

9.1 Pooled data: identical parameters

We could pool all of the data together. In other words, let's say we measure r_1 out of n_1 reversals in the first set of experiments, r_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

$$\begin{aligned} r &= \sum_{i=1}^{k} r_i \\ \text{out of } n &= \sum_{i=1}^{k} n_i \text{ reversals.} \end{aligned} \tag{9.2}$$

We then compute our posterior as in equation (9.1). Here, the assumption is that the result in each experiment are governed by *identical parameters*. That is to say that we assume $p_1 = p_2 = \cdots = p_k = p$.

This is similar to what we did in section 1.9, in which we looked at how a single hypothesis (or parameter value) is informed by more data.

9.2 Independent parameters

As an alternative, we could instead say that the parameters in each experiment are totally independent of each other. In this case, we assume that $p_1, p_2, \ldots, p_k$ are all independent of each other. The likelihoods and priors all multiply and the posterior probability is

$$g(\mathbf{p} \mid \mathbf{r}, \mathbf{n}) = \prod_{i=1}^{k} \frac{(n_i + 1)!}{(n_i - r_i)! r_i!} p_i^{r_i} (1 - p_i)^{n_i - r_i}, \tag{9.3}$$

where $\mathbf{p} = \{p_1, p_2, \ldots p_k\}$, with $\mathbf{n}$ and $\mathbf{r}$ similarly defined, and the posterior is understood to be zero if any the p_i's fall out of the interval $[0, 1]$.

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

9.3 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 p 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 q, and the values of p_i may vary from this q according to some probability distribution, $g(p_i \mid q)$. So now, we have parameters $p_1, p_2, \ldots, p_k$ and q. So, the posterior can be written using Bayes's theorem,

$$g(q, \mathbf{p} \mid \mathbf{r}, \mathbf{n}) = \frac{f(\mathbf{r}, \mathbf{n} \mid q, \mathbf{p})\, g(q, \mathbf{p})}{f(\mathbf{n}, \mathbf{r})}. \tag{9.4}$$

Note, though, that the observed values of r do not depend directly on q, only on $\mathbf{p}$. In other words, the observations are only *indirectly* dependent on q. So, we can write $f(\mathbf{r}, \mathbf{n} \mid q, \mathbf{p}) = f(\mathbf{r}, \mathbf{n} \mid \mathbf{p})$. Thus, we have

$$g(q, \mathbf{p} \mid \mathbf{r}, \mathbf{n}) = \frac{f(\mathbf{r}, \mathbf{n} \mid \mathbf{p})\, g(q, \mathbf{p})}{f(\mathbf{n}, \mathbf{r})}. \tag{9.5}$$

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

$$g(q, \mathbf{p}) = g(\mathbf{p} \mid q)\, g(q). \tag{9.6}$$

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

$$g(q, \mathbf{p} \mid \mathbf{r}, \mathbf{n}) = \frac{f(\mathbf{r}, \mathbf{n} \mid \mathbf{p})\, g(\mathbf{p} \mid q)\, g(q)}{f(\mathbf{n}, \mathbf{r})}. \tag{9.7}$$

Now, if we read off the numerator of this equation, we see a chain of dependencies. The experimental results $\mathbf{r}$ depend on parameters $\mathbf{p}$. Parameters $\mathbf{p}$ depend on *hyperparameter* q. Hyperparameter q 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 $g(\mathbf{p} \mid q)\, g(q)$ comprises the prior, as it is independent of the observed data.

9.4 Exchangeability

The conditional probability, $g(\mathbf{p} \mid q)$, can take any reasonable form. In the case where we have no reason to believe that we can distinguish any one p_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., $g(p_1, p_2, \ldots, p_k \mid q)$ is invariant to permutations of the indices. Parameters behaving this way are said to be **exchangeable**. A common (simple) exchangeable distribution is

$$g(\mathbf{p} \mid q) = \prod_{i=1}^{k} g(p_i \mid q), \tag{9.8}$$

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

9.5 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, $g(p_i \mid q)$, as well as $g(q)$ For the latter, we will take it to be uniform on $[0, 1]$. This is equivalent to taking it to be a Beta distribution with $\alpha = \beta = 1$. The Beta distribution is a good choice in this case, as it is a probability distribution of probabilities. For the conditional distribution $g(p_i \mid q)$, we also assume it is Beta-distributed.

The Beta distribution is typically written as

$$g(p \mid \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}, \tag{9.9}$$

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

$$q = \frac{\alpha}{\alpha+\beta}, \tag{9.10}$$

$$\kappa = \alpha + \beta. \tag{9.11}$$

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

Because the Beta distribution has two parameters, we cannot just parametrize the model with q. We would have to use q and κ or alternatively α and β. So, our expression for the posterior is

$$g(\alpha, \beta, \mathbf{p} \mid \mathbf{r}, \mathbf{n}) = \frac{f(\mathbf{r}, \mathbf{n} \mid \mathbf{p})\, g(\alpha, \beta) \prod_{i=1}^{k} g(p_i \mid \alpha, \beta)}{f(\mathbf{n}, \mathbf{r})}. \tag{9.12}$$

Alternatively, we could parametrize the model in terms of q and κ, giving

$$g(q, \kappa, \mathbf{p} \mid \mathbf{r}, \mathbf{n}) = \frac{f(\mathbf{r}, \mathbf{n} \mid \mathbf{p})\, g(q, \kappa) \prod_{i=1}^{k} g(p_i \mid q, \kappa)}{f(\mathbf{n}, \mathbf{r})}. \tag{9.13}$$

Note that if we do choose to parametrize our model with q and κ, we can convert back to α and β using

$$\alpha = q\kappa \tag{9.14}$$

$$\beta = (1-q)\kappa. \tag{9.15}$$

9.6 Choice of prior

As already stated, the likelihood is Binomial, with

$$f(\mathbf{r}, \mathbf{n} \mid \mathbf{p}) = \prod_{i=1}^{k} f(r_i, n_i \mid p_i) = \prod_{i=1}^{k} \frac{n_i!}{r_i!(n_i - r_i)!} p_i^{r_i}(1-p_i)^{n_i - r_i}, \tag{9.16}$$

and $g(p_i \mid \alpha, \beta)$ is Beta distributed, with

$$g(p_i \mid \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p_i^{\alpha-1}(1-p_i)^{\beta-1}. \tag{9.17}$$

We are now left to specify the hyperprior, $g(\alpha, \beta)$. We might choose to specify the prior in terms of q and κ, since these seem at face to be more intuitive. We can take a Uniform prior for q and a Jeffreys prior for κ, as we often do. That is, $g(q, \kappa) \propto 1/\kappa$. Applying the change of variables formula, we have

$$g(\alpha, \beta) \propto \left| \begin{matrix} \frac{\partial q}{\partial \alpha} & \frac{\partial q}{\partial \beta} \\ \frac{\partial \kappa}{\partial \alpha} & \frac{\partial \kappa}{\partial \beta} \end{matrix} \right| \frac{1}{\alpha + \beta} = \left| \begin{matrix} \frac{\beta}{(\alpha+\beta)^2} & -\frac{\alpha}{(\alpha+\beta)^2} \\ 1 & 1 \end{matrix} \right| \frac{1}{\alpha + \beta} = \frac{1}{(\alpha + \beta)^2}. \tag{9.18}$$

So, a uniform prior for q and a Jeffreys prior for κ results in a uniform prior in α and β, defined on $\alpha, \beta \in (0, \text{inf})$. If we use this Uniform prior, we have

$$\begin{aligned} g(\alpha, \beta, \mathbf{p} \mid \mathbf{r}, \mathbf{n}) &\propto \frac{1}{(\alpha + \beta)^2} \prod_{i=1}^{k} \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} p_i^{\alpha-1}(1 - p_i)^{\beta-1} p_i^{r_i} (1 - p_i)^{n_i - r_i} \\ &\propto \frac{1}{(\alpha + \beta)^2} \prod_{i=1}^{k} \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} p_i^{r_i+\alpha-1} (1 - p_i)^{n_i - r_i + \beta - 1}. \end{aligned} \tag{9.19}$$

We can integrate the right hand side over $p_1, p_2, \ldots$ to get the marginalized posterior for the hyperparameters α and β. We can do the integral by inspection, noting that $p_i^{r_i+\alpha-1}(1 - p_i)^{n_i - r_i + \beta - 1}$ is the same functional form of an unnormalized Beta distribution, so we must have

$$\int_0^1 \mathrm{d}p_i \, p_i^{r_i+\alpha-1}(1 - p_i)^{n_i - r_i + \beta - 1} = \frac{\Gamma(r_i + \alpha)\Gamma(n_i - r_i + \beta)}{\Gamma(n_i + \alpha + \beta)}. \tag{9.20}$$

So, the unnormalized marginalized posterior is

$$g(\alpha, \beta \mid \mathbf{r}, \mathbf{n}) \propto \frac{1}{(\alpha + \beta)^2} \prod_{i=1}^{k} \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(r_i + \alpha)\Gamma(n_i - r_i + \beta)}{\Gamma(n_i + \alpha + \beta)}. \tag{9.21}$$

There is a problem with this posterior: it is improper. That is to say that it is unnormalizable. This can be seen by using the reciprocal relation for gamma functions, $x\Gamma(x) = \Gamma(x + 1)$ to re-write the marginalized posterior.

$$\begin{aligned} g(\alpha, \beta \mid \mathbf{r}, \mathbf{n}) &\propto \frac{1}{(\alpha + \beta)^2} \prod_{i=1}^{k} \frac{\left(\prod_{m=0}^{r_i-1}(\alpha + m)\right)\left(\prod_{m=0}^{n_i-r_i-1}(\beta + m)\right)}{\prod_{m=0}^{n-1}(\alpha + \beta + m)} \\ &= \frac{1}{(\alpha + \beta)^2} \prod_{i=1}^{k} \frac{\mathcal{O}(\alpha^{r_i})\,\mathcal{O}(\beta^{n_i-r_i})}{\mathcal{O}\left((\alpha + \beta)^{n_i}\right)} \end{aligned}$$

$$= \frac{1}{(\alpha+\beta)^2} \prod_{i=1}^{k} \mathcal{O}\left(\left(\frac{\alpha}{\alpha+\beta}\right)^{r_i}\right) \mathcal{O}\left(\left(\frac{\beta}{\alpha+\beta}\right)^{n_i-r_i}\right). \tag{9.22}$$

Since we have $q = \alpha/(\alpha+\beta) = (1+\beta/\alpha)^{-1}$ and must lie between zero and one, we can consider a limit of large α and β with the ratio α/β fixed at some constant, finite value. Then, for large α and β, the product term in the expression for the unnormalized marginal posterior is constant. Therefore, the integral

$$\int_0^\infty \mathrm{d}\alpha \int_0^\infty \mathrm{d}\beta \; g(\alpha, \beta \mid \mathbf{r}, \mathbf{n}) \tag{9.23}$$

diverges because the integral over $(\alpha+\beta)^{-2}$ diverges. This gives an improper *posterior*, which is not acceptable.

It turns out that this problem occurs generally in hierarchical models. The variance of a Beta distribution is approximately proportional to κ^{-1}, especially at large α and β. By choosing a Jeffreys prior for the variance, we are choosing a Uniform prior for the log of the variance. When we do this with hierarchical models, that is choose a Uniform prior for the log of the variance of a hyperprior for exchangeable parameters, we get an improper posterior.

So, it is often tricky to be truly uninformative with your hyperpriors. For the present example, we will instead choose a Uniform prior in the standard deviation, so that $\kappa^{-1/2}$ has a Uniform prior; $g(q, \kappa^{-1/2}) = \text{constant}$. If we do this, we have

$$g(\alpha, \beta) \propto \begin{vmatrix} \frac{\partial q}{\partial \alpha} & \frac{\partial q}{\partial \beta} \\ \frac{\partial \sqrt{\kappa}}{\partial \alpha} & \frac{\partial \sqrt{\kappa}}{\partial \beta} \end{vmatrix} = \begin{vmatrix} \frac{\beta}{(\alpha+\beta)^2} & -\frac{\alpha}{(\alpha+\beta)^2} \\ -\frac{1}{2(\alpha+\beta)^{3/2}} & -\frac{1}{2(\alpha+\beta)^{3/2}} \end{vmatrix} \propto \frac{1}{(\alpha+\beta)^{5/2}}. \tag{9.24}$$

With this prior, we have an unnormalized posterior of

$$g(\alpha, \beta, \mathbf{p} \mid \mathbf{r}, \mathbf{n}) \propto \frac{1}{(\alpha+\beta)^{5/2}} \prod_{i=1}^{k} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p_i^{r_i+\alpha-1} (1-p_i)^{n_i-r_i+\beta-1}. \tag{9.25}$$

This is a proper posterior, which you can prove with similar arguments as we made to show that the first posterior we considered was *im*proper.

9.7 Implementation

In some cases, we can do some gnarly integration and work out analytical results for the posterior of a hierarchical model. This usually involves choosing conjugate priors. Most often, though, we need to resort to numerical methods, MCMC as usual being the most powerful. To see the worm reversal problem solved with a hierarchical model, see the implementation here.

9.8 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

$$g(\theta, \phi \mid D) = \frac{f(D \mid \theta)\, g(\theta \mid \phi)\, P(\phi)}{f(D)} \tag{9.26}$$

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

$$g(\theta, \phi, \xi \mid D) = \frac{f(D \mid \theta)\, g(\theta \mid \phi)\, g(\phi \mid \xi)\, g(\xi)}{f(D)}, \tag{9.27}$$

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 $f(D \mid \theta)$, 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.