1 Hierarchical Models

We will closely follow the back half of lecture 17 (that we did not discuss Tuesday) for this portion of the lab. The idea here is that, rather than formulating a prior from thin air, suppose that we have what we can regard as replications of an experiment (rather than just observing a single experiment like we have done previously). This idea of replications is an assumption on our data called *exchangability*. The idea behind hierarchical models is that we can use this collection of observations to actually **learn the prior**. This feels much safer than having to specify a prior as we have been doing. Of course, since we are Bayesians, we have to specify distributions on the things we are to learn. These sit on a different level (above the prior) and are thus called **hyperpriors** that describe our beliefs on **hyperparameters**, a term that we have used in the past (the parameters of our prior). This new idea may lead us to think that if we don't know how to put hyperpriors on our hyperparameters, we should put hyper-hyper priors on these, and then learn those. This sounds like we are moving towards a "turtles all the way down" scenario. The good news is that the further down we bury the actual fixed objects that we need to choose, the less they affect the outcome. So we are just going to stop for now with hyperpriors.

We'll look at two situations: normal data and binomial data. We will use pre-existing code for both, but first we are going to argue that the Gibbs sampler is going to be available to us and quite convenient.

2 Normal data

We are going to observe what we will denote as $\bar{x}_i$. These values we will assume were generated from a normal distribution with mean θ_i and standard deviation $\frac{\sigma}{\sqrt{n}}$. In lab 3, we assumed that θ_i came from a normal distribution with mean μ and variance $\frac{1}{\tau}$. Then we chose these values. Here, we will learn them, so we will then put a prior on μ and τ.

a. There are a handful of ingredients here: a hyper-prior on μ, a hyper-prior on τ, a prior on each of the θs, and then a distribution on each of the $\bar{x}$ values. Draw a graph with each of these as a vertex with an edge connecting any two vertices if the behavior of one explicitly depends on another.

b. What is it about the structure of this graph that makes a Gibbs sampler implementation so appealing?

c. Run the code from lecture 17 on the SAT preparation exam and discuss how the estimates differ from the frequentist estimators that would either i) assume all groups are unique, or ii) assume that all groups are estimating the same

quantity.

d. Simulate data from the model (choosing only the values at the highest level) and see how this performs in estimating the quantities of interest (the θ_i values). The code is simplified in that it assumes that all you see is the $\bar{x}_i$ and have access to $\frac{\sigma^2}{n}$, which will both be inputted. If you had access to the raw data, then you could learn σ^2 as well; recall lab 3 to realize that that's going to be a lot messier.

3 Binomial data

Here, we will observe the outcome of several non-identical but conditionally independent binomial experiments, so we will observe x_i and n_i coming from a $\text{Bin}(n_i, p_i)$ experiment for $i = 1, \ldots, k$. We will assume that these p_i all came from a common $\text{Beta}(\alpha, \beta)$ distribution, and our task is to learn the value of these hyper-parameters (rather than just fixing them as we have done early on).

An example is that we have some new treatment for some oft-terminal disease. We implement the treatment in k different hospitals, and see how many people survive (x_i out of n_i) in each. We think there are likely to be differences between these hospitals (the doctors are different, the patient population is different, etc). For instance, Loma Linda, about 25 miles east of Claremont, is one of five "Blue Zones" (`https://en.wikipedia.org/wiki/Blue_Zone`) globally, where one might expect a higher proportion of survivors if LLUMC was chosen as one of the hospitals. Of course, if we knew one was Loma Linda, knowing what we know, we could not regard this data as *exchangable* any more, and would need to model in this information, making the model quite a bit more cumbersome.

The posterior conditionals for the Gibbs sampler are all easy (standard calculations) except for the conditional on α and β, the hyper-parameters. These are not of recognizable form, and the code we will use relies on a Metropolis-Hastings step to generate these as we move through the Gibbs sampler (which you can convince yourself will still lead to the right stationary distribution, but could affect the convergence properties). The code is available here: `http://www.stat.cmu.edu/~brian/724/week06/lec15.r` with the corresponding lecture notes here: `http://www.stat.cmu.edu/~brian/724/week06/lec15-mcmc2.pdf` if you want to see the calculations.

e. Rather than use their rat data (similar to the example above), use a different data set where you might actually know the right answer (or something like it). My idea is this: we are 2 weeks into the 2018 MLB season, so hitters have fairly unstable batting averages currently. Their batting averages at the end of last season were much more stable (based on very large sample sizes). Let's try to guess their last season batting averages (which we will treat as their "true" batting average) based on this year's results. For instance, currently

Mike Trout is 12 for 51 (BA of .235). Last year, he hit .306. I know this from `https://sports.yahoo.com/mlb/players/8861/`. Grab a random sample of k "binomial" experiments, and then try to estimate the true binomial proportions. You don't need to use my baseball idea. But it's the only thing I could come up with.

i) Compare this via MSE to the two frequentist extremes. A plot showing the truth, the data, and the estimate would be cool.

ii) Convince me that you have monitored the convergence of your chain.

iii) Also, recall from Tuesday that David Robinson estimated $\alpha = 78.66$ and $\beta = 224.87$ in his article here: `http://varianceexplained.org/r/empirical_bayes_baseball/`. Does your much smaller data set suggest that these values are reasonable?

A note: The above article is doing something called *empirical bayes*. It's using frequentist theory (MLEs) to find the hyper-parameters on the prior, and then the rest of the analysis proceeds as if these were known constants. So this uncertainty isn't modeled in and isn't reflected in your posterior analysis. Of course, with his sample size, this won't make much of a difference. Hierarchical models are fully Bayesian versions of this idea, and will be much more accurate when this uncertainty is non-negligible. The trade-off, of course, is that the calculations he has to do after choosing α and β are easy (the one's we've been doing all semester).