5: Hierarchical models

09/23/19

Introduction

On the one hand, we can assume that our data are iid, conditional on one parameter. On the other hand, we can assume that each data point gets its own parameter. The former might be too inflexible, while the latter might be too flexible, leading too overfitting.

Hierarchical models are a good "in between" option that allows each data point to get its own parameter; however, these parameters are "tied together" in a certain sense.

Rat tumor example

1. $j = 1, 2, \ldots, 71$ groups/experiments
2. θ_j is the probability of any rat getting a tumor in experiment j
3. θ_j are all different because of rat and/or experimental differences
4. y_j is the count of rats with tumors in experiment j (out of n_j total rats)
5. $y_j \mid \theta_j, n_j \sim \text{Binomial}(n_j, \theta_j)$ exchangeable
6. $\theta_j \overset{\text{iid}}{\sim} \text{Beta}(\alpha, \beta)$

Rat tumor example

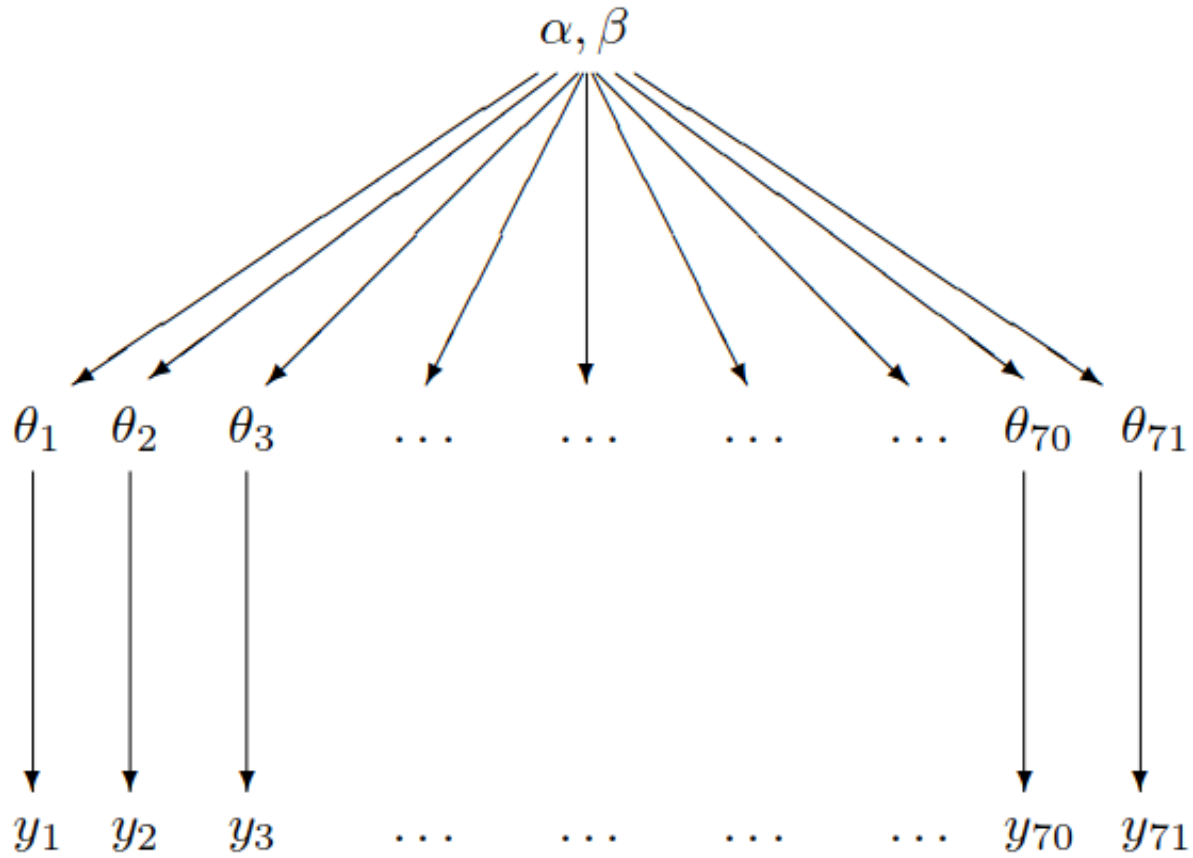

Figure 5.1: *Structure of the hierarchical model for the rat tumor example.*

Rat tumor example

Consider groups $1, \ldots, 70$ as historical data. We are interested specifically in θ_{71}.

Naive approach: only choose Beta(α, β) prior for θ_{71}. Choose α, β based on historical data $y_1, \ldots, y_{71}$, but in an ad hoc way, by setting the prior mean to be the empirical mean, and the prior variance equal to the sample variance. I.e. by solving

$$\begin{bmatrix} \hat{p} = 70^{-1} \sum_{i=1}^{70} y_j/n_j \\ 70^{-1} \sum_{i=1}^{70} (y_j/n_j - \hat{p})^2 \end{bmatrix} = \begin{bmatrix} \alpha/(\alpha+\beta) \\ \alpha\beta/\{(\alpha+\beta)^2(\alpha+\beta+1)\} \end{bmatrix}$$

You end up with $(\alpha, \beta) = (1.4, 8.6)$. Then, because $(y_{71}, n_{71}) = (4, 14)$,

$$\theta_{71} \mid y_{71} \sim \text{Beta}(5.4, 18.6)$$

Rat tumor example

The above approach is called **empirical Bayes**. Problems with this approach:

1. can't really make inferences on $\theta_1, \ldots, \theta_{70}$ unless you "use the data twice"
2. how do we know we used the right point estimates for prior construction? The point estimates ignores uncertainty
3. Should α, β be estimated? They are part of prior distributions on θ_{71} and thus not necessarily be known.

Exchangeability in the prior

Why do we use exchangeable priors?

Q: If someone told you $\theta_1 = .2, \theta_2 = .3$, would you react differently than if they told you $\theta_2 = .2, \theta_1 = .3$?

Exchangeability in the prior

Why do we use exchangeable priors?

Q: If someone told you $\theta_1 = .2, \theta_2 = .3$, would you react differently than if they told you $\theta_2 = .2, \theta_1 = .3$?

A1: "No, I don't know anything about these labs, so it's all the same to me."
This means $p(\theta_{1:71})$ should be chosen to be exchangeable.

A2: "Yes, the second one is rarer a priori. I think θ_1 should be higher because the first lab sources their rats from NYC subways, and the second sources theirs from DC subways."

This means $p(\theta_{1:71})$ should not be chosen to be exchangeable.

Exchangeability in the prior

For many examples, no information other than the observations y is available to distinguish any of the θ_j from any of the others for all $j = 1, \ldots, J$.

Assume exchangeability in y. It is natural to assume exchangeability in θ.

Exchangeability in the prior

Let's say we assume exchangeability. How can we pick a prior?

Option 1: iid (not a hierarchical model)

$$p(\theta_{1:J}) = \prod_{i=1}^{J} p(\theta_i)$$

Does your opinion about θ_1 change if we knew θ_2? If yes, this isn't appropriate.

Exchangeability in the prior

Option 2: Simplest form of hierarchical exchangeable distribution

$$p(\theta \mid \phi) = \prod_{j=1}^{J} p(\theta_j \mid \phi)$$

In general ϕ is unknown, we assume $\phi \sim p(\phi)$

$$p(\theta) = \int \left(\prod_{j=1}^{J} p(\theta_j \mid \phi) \right) p(\phi) d\phi$$

E.g. θ_i and θ_j are positively correlated.

Exchangeability in the prior

If observations are not fully exchangeable, but are *partially* or *conditionally* exchangeable:

- Partially exchangeable: observations are grouped, group properties θ_i are assumed to be exchangeable
- Conditionally exchangeable: additional information is passed along with y_i, say x_i, (y_i, x_i) are exchangeable
  θ_i: a joint model for (y_i, x_i) or a conditional model for $y_i \mid x_i$

Option 3:

$$p(\theta \mid x) = \int \left(\prod_{j=1}^{J} p(\theta_j \mid \phi, x_i) \right) p(\phi \mid x) d\phi$$

$x = (x_1, \ldots, x_J)$
In rat tumor example, n_i can be treated as x_i and prior dependence is dropped because there is no indication that y_i/n_i is closely related to n_i

Rat tumor example – A full Bayesian approach

A "better" way:

1. choose (hyper)prior $p(\alpha, \beta)$
2. choose prior $p(\theta_{1:71} \mid \alpha, \beta)$
3. choose likelihood $p(y \mid \theta_{1:71}, \alpha, \beta) = p(y \mid \theta_{1:71}) = \prod_{j=1}^{71} p(y_j \mid \theta_j)$

Then

$$\begin{aligned} p(\theta_{1:71}, \alpha, \beta \mid y) &\propto p(y \mid \theta_{1:71}, \alpha, \beta) p(\theta_{1:71} \mid \alpha, \beta) p(\alpha, \beta) && \text{(Bayes')} \\ &= p(y \mid \theta_{1:71}) p(\theta_{1:71} \mid \alpha, \beta) p(\alpha, \beta) && \text{(condtl. indep.)} \end{aligned}$$

Question: how do we draw simulations from the above posterior distribution?

Posterior Predictive Distributions: two choices

If you want the predictive distribution of new rat counts ($\tilde{y}_{54}$) in an old/existing experiment (say $j = 54$), then you can use

$$p(\tilde{y}_{54} \mid y) = \int p(\tilde{y} \mid \theta_{54}) p(\theta_{54} \mid y) x \mathrm{d}\theta_{54}$$

If you want the probability distribution of future rat counts ($\tilde{y}_{72}$) in a future experiment (say $j = 72$) coming from the same "superpopulation", you can use

$$p(\tilde{y}_{72} \mid y) = \iiint p(\tilde{y}_{72} \mid \theta_{72}) p(\theta_{72} \mid \alpha, \beta) p(\alpha, \beta \mid y) \mathrm{d}\theta_{72} \mathrm{d}\alpha \mathrm{d}\beta$$

Both strategies are based on the same decomposition, but the second way simulates twice.