Intro to Hierarchical Models

ISTA 410 / INFO 510: Bayesian Modeling and Inference

U. of Arizona School of Information

February 22, 2021

Outline

Last week:

- Posterior predictive checking
- Simple linear regression; the role of priors
- Specifying and sampling from a model in PyMC3

This week: hierarchical/multilevel models

Intro to hierarchical models

How much traffic is bicycle traffic?

In the book there is a data set with observations of bicycle traffic on a number of streets.

- Data collected by standing at the roadside for some amount of time
- Count number of bicycles and number of non-bicycle vehicles
- Parameter of interest: proportion of traffic that is bicycles

Fully pooled model

A fully pooled model:

$$y_j \sim \text{Binomial}(\theta, n_j)$$
$$\theta \sim \text{Beta}(\alpha_0, \beta_0)$$

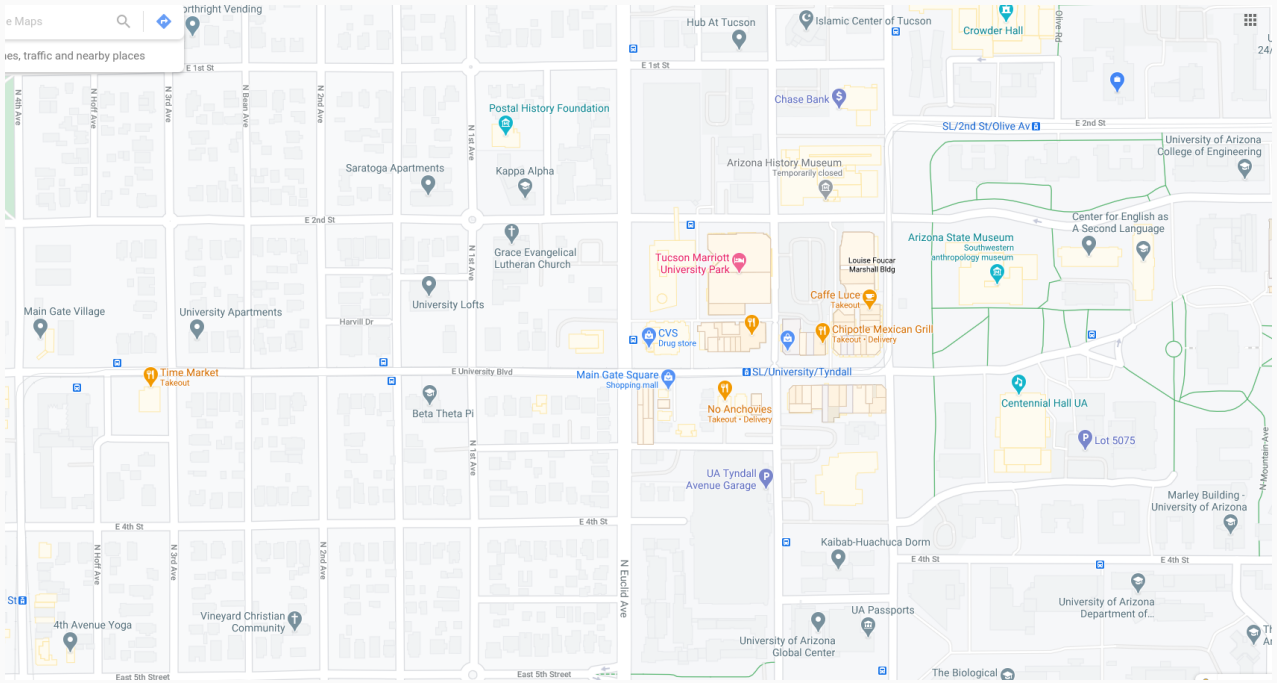
for fixed α_0, β_0.

- Choosing $\alpha_0 = 1, \beta_0 = 1$ gives a completely noninformative (flat) prior
- Weakly informative prior also reasonable, e.g. $\alpha_0 = 1, \beta_0 = 3$ for prior mean of 25% bicycle traffic

Why not pool?

This model we wrote in the previous section treats all streets as the same; each street's observation is an observation of the same underlying proportion.

But this isn't particularly reasonable:



Why multilevel modeling?

Imagine you're collecting data for this.

- You go out to University Blvd/1st Ave and count cars and bicycles for 30 minutes.
- Is this count going to be representative of the level of bicycle traffic in Tucson?

Why multilevel modeling?

Imagine you're collecting data for this.

- You go out to University Blvd/1st Ave and count cars and bicycles for 30 minutes.
- Is this count going to be representative of the level of bicycle traffic in Tucson?
    - No; bike traffic is much higher on University than other streets

Why a hierarchical model?

As an alternative, we could treat each street as an independent entity:

Why a hierarchical model?

As an alternative, we could treat each street as an independent entity:

$$y_j \sim \text{Binomial}(\theta_j, n_j)$$
$$\theta_j \sim \text{Beta}(\alpha_0, \beta_0)$$

for fixed α_0, β_0.

- Exactly like the previous model, except we now have 10 independent θ_js for the 10 streets
- Same considerations for choice of prior

Call this the separate-effects model.

Comparing the two models

Let's jump over to a notebook and compare the results of these two models.

Why a hierarchical model?

Choosing between the two models: classically, do an analysis of variance

- Compare variance within groups (streets) to variance between streets
- Test against the null hypothesis that all streets are the same
- If we reject the null, take the separate-effects model
- If we don't take the pooled model

Problem: false dichotomy!

Why a hierarchical model?

In reality, it is most plausible that both of the following are true:

- The streets are not identical; some of the streets are more popular with cyclists
- Observations of one street can inform our knowledge of the others
- The high bicycle traffic we observe on University is:
  - partly a reflection of that street's individual geographical properties
  - partly a reflection of the city's relatively high bicycle friendliness

So: neither side of this dichotomy is preferable.

Variability at two levels

There is variability at two levels in this problem:

- Different cities have, overall, more or less bicycle traffic
- Within a given city, different streets have more or less bicycle traffic

So, observations of one street should inform our estimates of other streets, *without* making the same estimate for every street

The hierarchical model

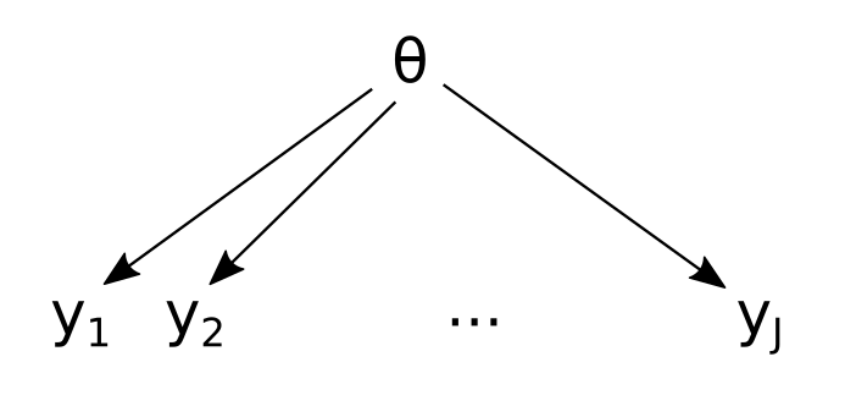
With a Bayesian approach, we can find a compromise.

- We have a θ for each street
- However, instead of being fully independent, each θ is drawn from a common probability distribution
- This probability distribution, a *hyperprior*, depends on *hyperparameters* which we estimate from the data

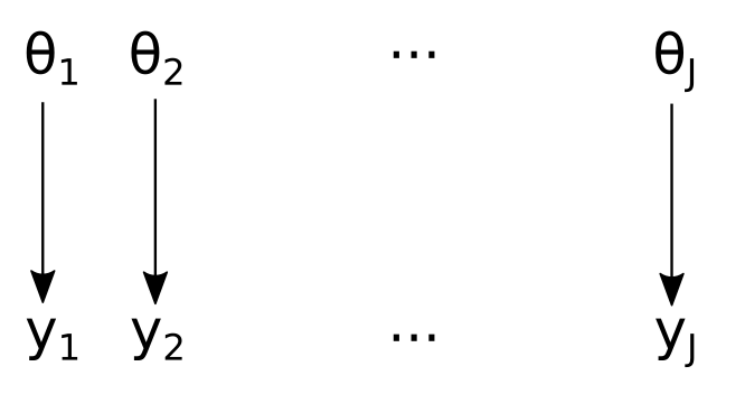
(note: slightly different sense of the term *hyperparameter* from its common use in ML)

Examining this graphically

Pooled model:

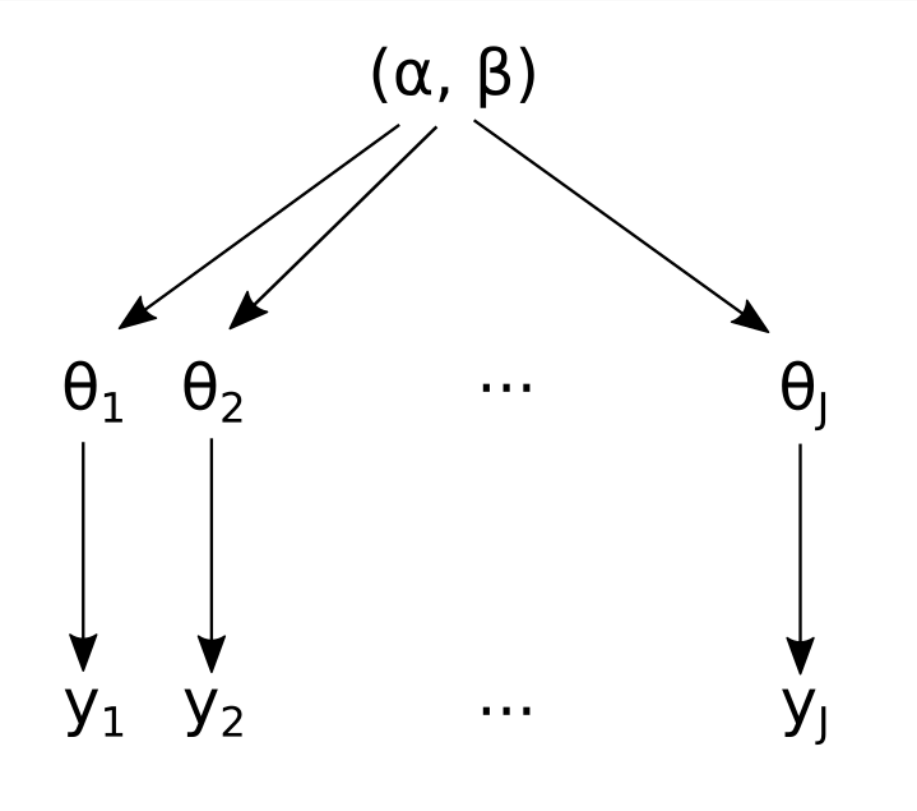


Separate model:



Examining this graphically

Hierarchical model combines the features of these two:



Setting up the model

This is conceptually only a slight difference from our previous model:

$$
\begin{aligned}
y_j &\sim \text{Binomial}(\theta_j, n_j) \\
\theta_j &\sim \text{Beta}(\alpha, \beta) \\
p(\alpha, \beta) &\propto ???
\end{aligned}
$$

Choosing a hyperprior

We need a prior distribution for α, β; this can be a tricky part of this sort of modeling, because the interpretation of these parameters is not so simple compared to θ_j.

Starting with the idea that α and β can represent "pseudocounts", parameterize in terms of

- $\mu = \frac{\alpha}{\alpha+\beta}$ – prior expectation
- $\eta = \alpha + \beta$ – prior "sample size" (think of this like a precision)

Choosing a hyperprior

Prior for μ:

- Need $0 < \mu < 1$, so we'll choose a Beta
- Informative version: cars outnumber bikes, so try $\mathrm{Beta}(1, 3)$

Prior for η:

- $\eta > 0$, so choose something with range $(0, \infty)$
- Fairly spread out, but put more prior mass near 0; try a half-Cauchy

Setting up the model

This is conceptually only a slight difference from our previous model:

$$
\begin{aligned}
y_j &\sim \text{Binomial}(\theta_j, n_j) \\
\theta_j &\sim \text{Beta}(\alpha, \beta) \\
\mu &:= \frac{\alpha}{\alpha + \beta} \\
\eta &:= \alpha + \beta \\
p(\mu) &\sim \text{Beta}(1, 3) \\
p(\eta) &\sim \text{HalfCauchy}(1)
\end{aligned}
$$

Note: BDA uses a vaguely similar but fairly opaque approach to reach a prior that is qualitatively fairly similar (details at the end of these slides)

Inference the hard way

As usual, we can make inferences by sampling from the posterior distribution. This can be done the hard way (directly), or the easy way (MCMC).

Hard way:

1. Calculate the posterior density $p(\alpha, \beta|y)$ on a grid of α and β values.
2. Sum over the β values to get an estimate of the marginal posterior $p(\alpha|y)$; use this to draw samples of α.
3. For each sampled value of α, use the conditional posterior $p(\beta|\alpha, y)$ (which is a slice of)
4. For each sampled pair (α_i, β_i), draw values of θ_j from the beta distribution $\mathrm{Beta}(\alpha_i + y_j, \beta_i + y_j - n_j)$

Inference the easy way

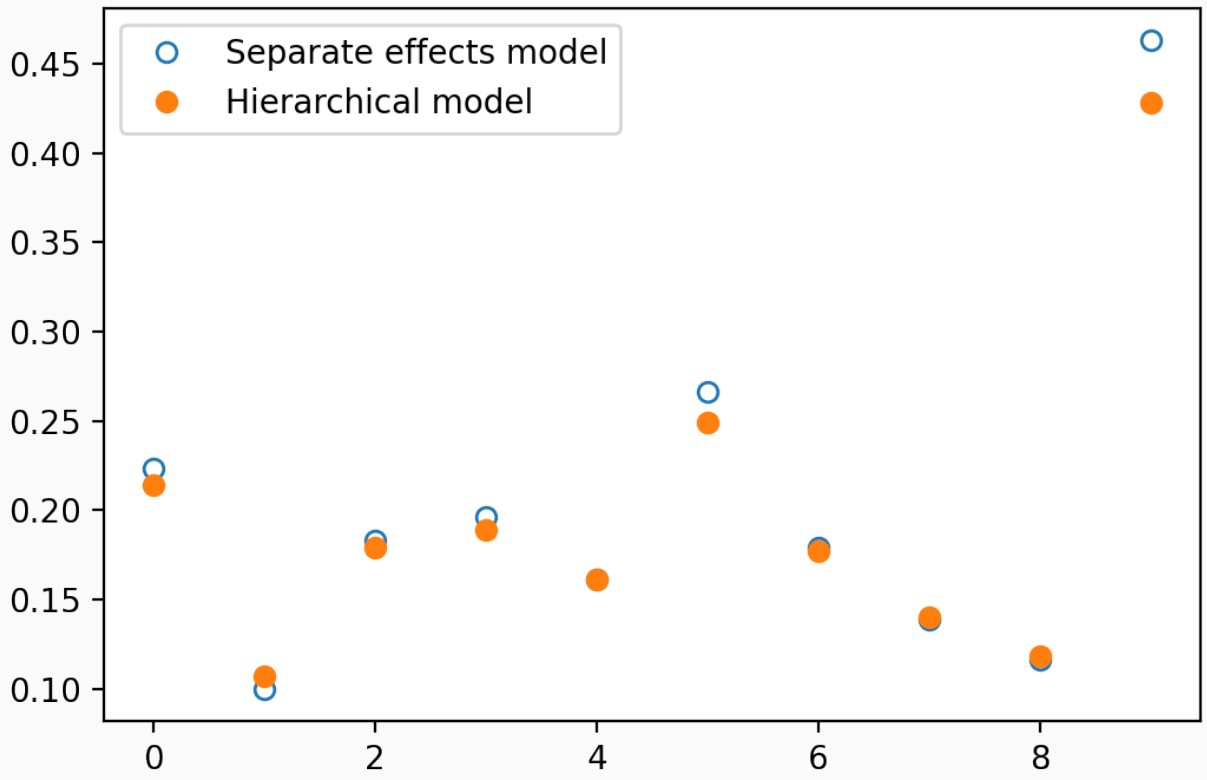
The easier approach: use MCMC to sample from the posterior.

Let's see this in action...

Comparison

What is the difference in the results?

Let's compare point estimates:



Shrinkage and regularization

The shrinkage effect we see is a form of regularization:

- Most extreme observations "shrunk" toward a central value
- Amount of shrinkage tuned to relative sample size

Difference: we learned the strength of regularization from the data

Underfitting and overfitting

Another way to think about this, in terms of underfitting and overfitting:

- The pooled model: strong underfitting
- The separate-effects model: strong overfitting
- Hierarchical model: adaptive regularization

With enough observations the seperate effects model will estimate each street similarly to the hierarchical model.

Independence and exchangeability

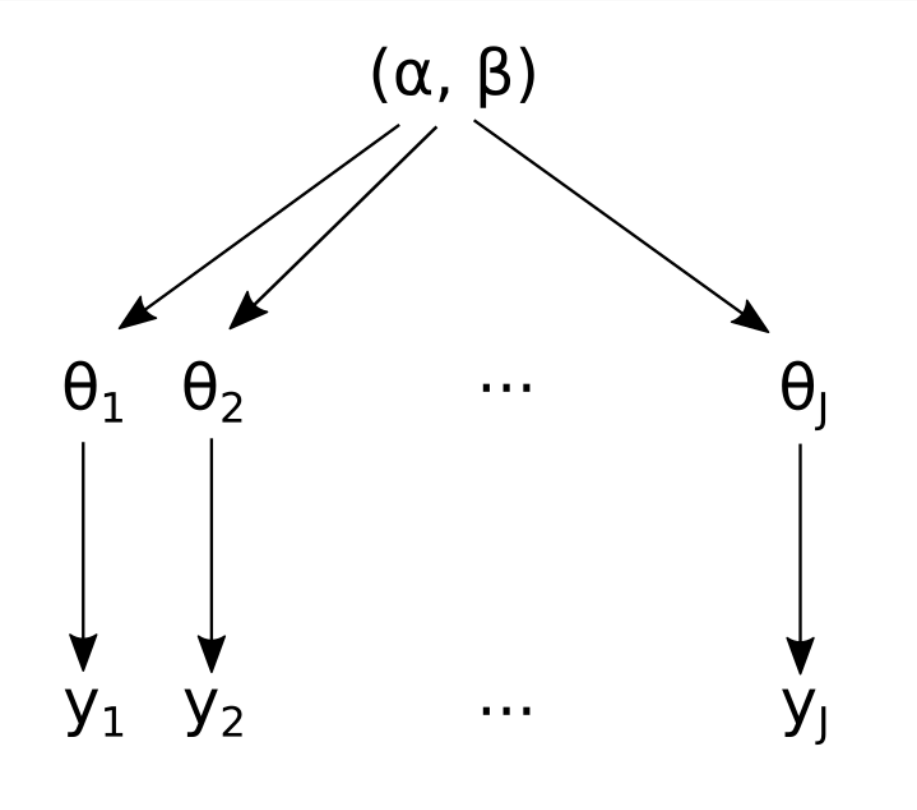
Independence of the θ_js

It's worth taking a moment to consider the independence properties of the parameters θ_j:

- In the hierarchical model, θ_js are not independent (they're independent in the pooled model)
- However, they satisfy two weaker properties:
  - conditional independence
  - exchangeability

Conditional independence

The θ_js are not independent, but *given fixed* α, β, they are:



Exchangeability

A closely related concept is *exchangeability*, which justifies the use of the hierarchical model:

- Observations are *exchangeable* if the joint probability distribution is invariant to permutations of the index
- Roughly: we would have the same model if we relabeled the $y_1, y_2, \ldots$
- Exchangeability is also evident in the directed graph model

Levels of exchangeability

The full data set contains observations from a total of 58 streets:

- small residential streets, medium streets, and busy arterial streets
- streets with or without bike lanes

Evidently, if we label the streets $y_1, \ldots, y_{58}$, they are not exchangeable.

But within the traffic/lane groups, the streets can be treated as exchangeable:

- Hierarchical model with several "levels"

Summary

Hierarchical models:

- Have several "levels" of parameters stacked
- Perform adaptive regularization – learn priors from the data

Next time: hierarchical normal model and computational difficulties

Appendix: hyperprior calculation

Choosing a hyperprior

BDA suggests the following as a hyperprior:

$$p(\alpha, \beta) \propto (\alpha + \beta)^{-5/2}$$

What's the intuition behind this?

Choosing a hyperprior

BDA suggests the following as a hyperprior:

$$p(\alpha, \beta) \propto (\alpha + \beta)^{-5/2}$$

What's the intuition behind this?

In a beta distribution, interpretation of parameters as "pseudocounts":

Choosing a hyperprior

BDA suggests the following as a hyperprior:

$$p(\alpha, \beta) \propto (\alpha + \beta)^{-5/2}$$

What's the intuition behind this?

In a beta distribution, interpretation of parameters as "pseudocounts":

- If we start with $\mathrm{Beta}(\alpha, \beta)$ and make binomial observations, we update to the posterior $\mathrm{Beta}(\alpha + n_s, \beta + n_f)$, with n_s successes and n_f failures
- So, we can think of α and β as "counts" of imaginary observations

Choosing a hyperprior

Goal: prior is noninformative on the mean value of θ_j and the spread, or scale, of that mean

- Mean is $\frac{\alpha}{\alpha+\beta}$
- Scale parameters (standard errors) for means are distributed like $n^{-1/2}$ where n is the sample size

So: set up a prior distribution that is uniform on $\left(\frac{\alpha}{\alpha+\beta}, (\alpha+\beta)^{-1/2}\right)$

Choosing a hyperprior

Define:

$$w = \frac{\alpha}{\alpha + \beta}$$

$$z = (\alpha + \beta)^{-1/2}$$

$$p(w, z) \propto 1$$

To get to

$$p(\alpha, \beta) \propto (\alpha + \beta)^{-5/2}$$

we have to do some calculus (on the following slides!)

Reminder

As a reminder, our prior distribution was uniform on

$$\left(\frac{\alpha}{\alpha+\beta}, (\alpha+\beta)^{-1/2}\right)$$

Define $w = \frac{\alpha}{\alpha+\beta}, z = (\alpha+\beta)^{-1/2}$, and set $p(w,z) \propto 1$.

Changing variables for probability densities comes from changing variables for integrals, because the PDF is defined by the property that

$$\mathrm{Pr}(\mathrm{x}_1, \ldots, \mathrm{x}_\mathrm{n} \in \mathrm{A}) = \int_\mathrm{A} \mathrm{p}(\mathrm{x}_1, \ldots, \mathrm{x}_\mathrm{n}) \mathrm{dx}_1 \ldots \mathrm{dx}_\mathrm{n}$$

To perform the change of variables, we need to multiply by the absolute determinant of the Jacobian matrix

$$J = \begin{pmatrix} \frac{\partial w}{\partial \alpha} & \frac{\partial w}{\partial \beta} \\ \frac{\partial z}{\partial \alpha} & \frac{\partial z}{\partial \alpha} \end{pmatrix}$$

To perform the change of variables, we need to multiply by the absolute determinant of the Jacobian matrix

$$J = \begin{pmatrix} \frac{\partial w}{\partial \alpha} & \frac{\partial w}{\partial \beta} \\ \frac{\partial z}{\partial \alpha} & \frac{\partial z}{\partial \alpha} \end{pmatrix}$$

$$J = \begin{pmatrix} \frac{\beta}{(\alpha+\beta)^2} & \frac{-\alpha}{(\alpha+\beta)^2} \\ -\frac{1}{2}(\alpha+\beta)^{-3/2} & -\frac{1}{2}(\alpha+\beta)^{-3/2} \end{pmatrix}$$

so $|\det J| = \frac{1}{2}(\alpha+\beta)^{-5/2}$ (and we can drop the $1/2$ because it's a constant)