Exchangeability and more hierarchical models

ISTA 410 / INFO 510: Bayesian Modeling and Inference

U. of Arizona School of Information

February 24, 2021

Outline

Last time:

- Bike lane example
- Hierarchical models
- Hyperprior selection

Now:

- Concept: exchangeability
- Hierarchical normal model

Recap

Bicycle traffic on neighborhood streets

Example from last time:

- Exercise 3.8 (and 5.13) in the textbook
- Data: observations of numbers of bicycles and other vehicles on neighborhood streets in Berkeley, CA
- Includes three classes of streets, with and without bike lanes
- We focus on one category: small streets with bike lanes

Goal: estimate the proportion of bicycle traffic

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

Fully separated model

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

Fully separated 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.

Setting up the model

A compromise: hierarchical model

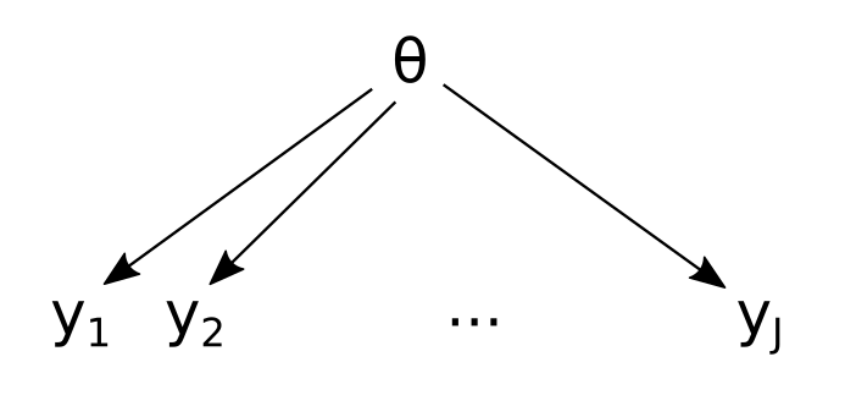
$$
\begin{aligned}
y_j &\sim \text{Binomial}(\theta_j, n_j) \\
\theta_j &\sim \text{Beta}(\alpha, \beta) \\
\mu &:= \frac{\alpha}{\alpha + \beta} \\
\eta &:= \alpha + \beta \\
p(\mu) &\sim \text{Beta}(1, 3) \\
p(\eta) &\sim \text{HalfCauchy}(1)
\end{aligned}
$$

Note: the book uses:

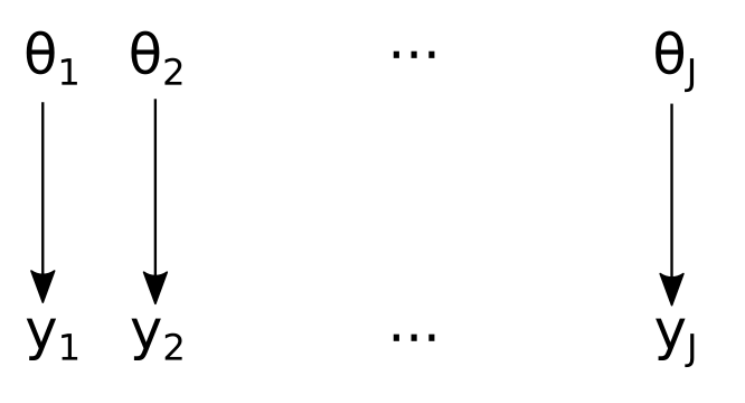
$$p(\alpha, \beta) \propto (\alpha + \beta)^{-5/2}$$

Examining this graphically

Pooled model:

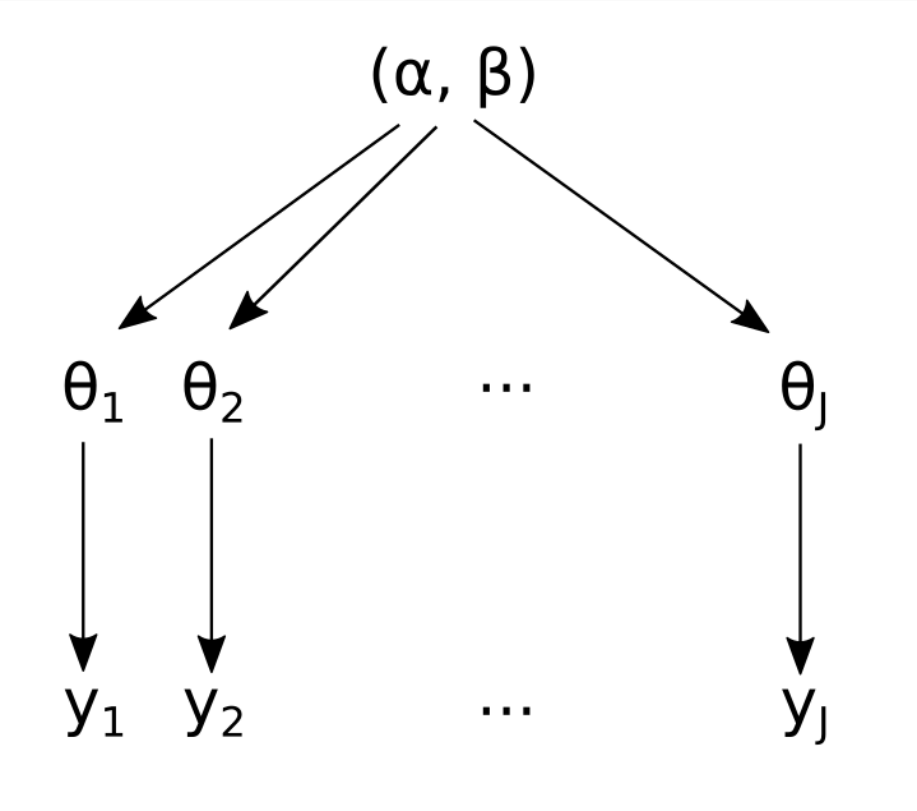


Separate model:



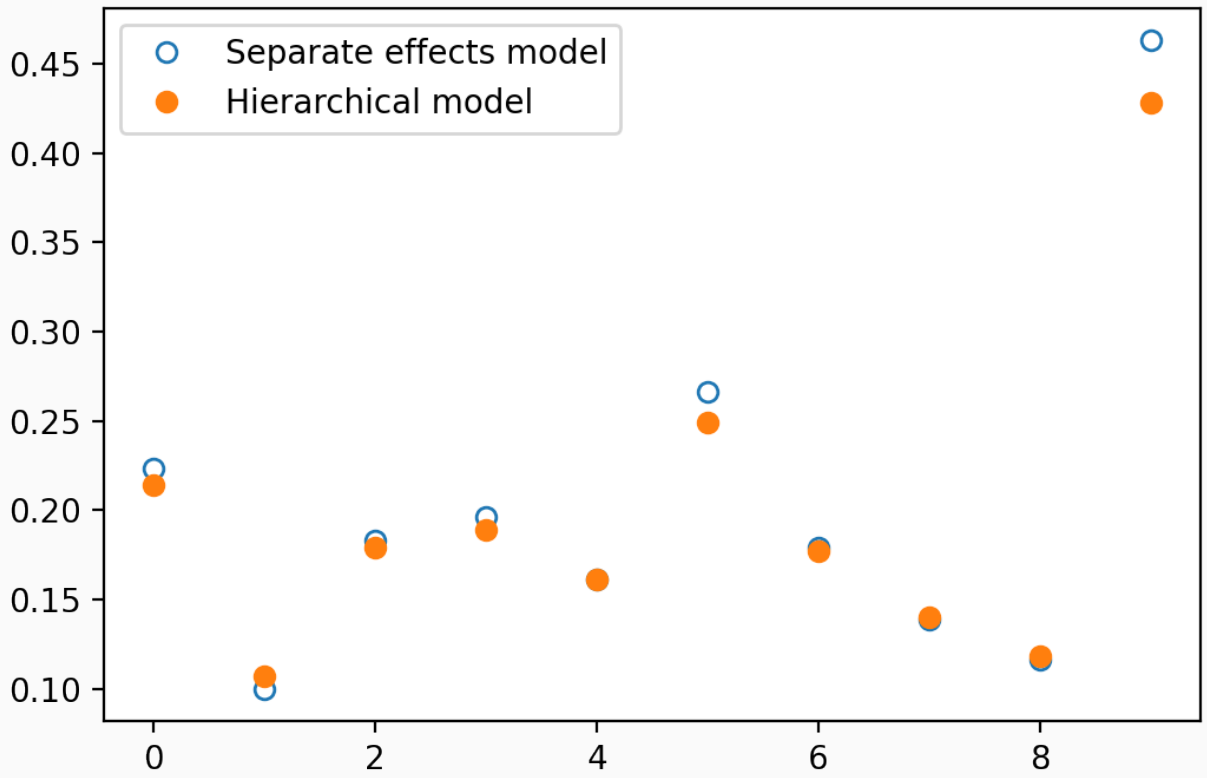
Examining this graphically

Hierarchical model combines the features of these two:



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

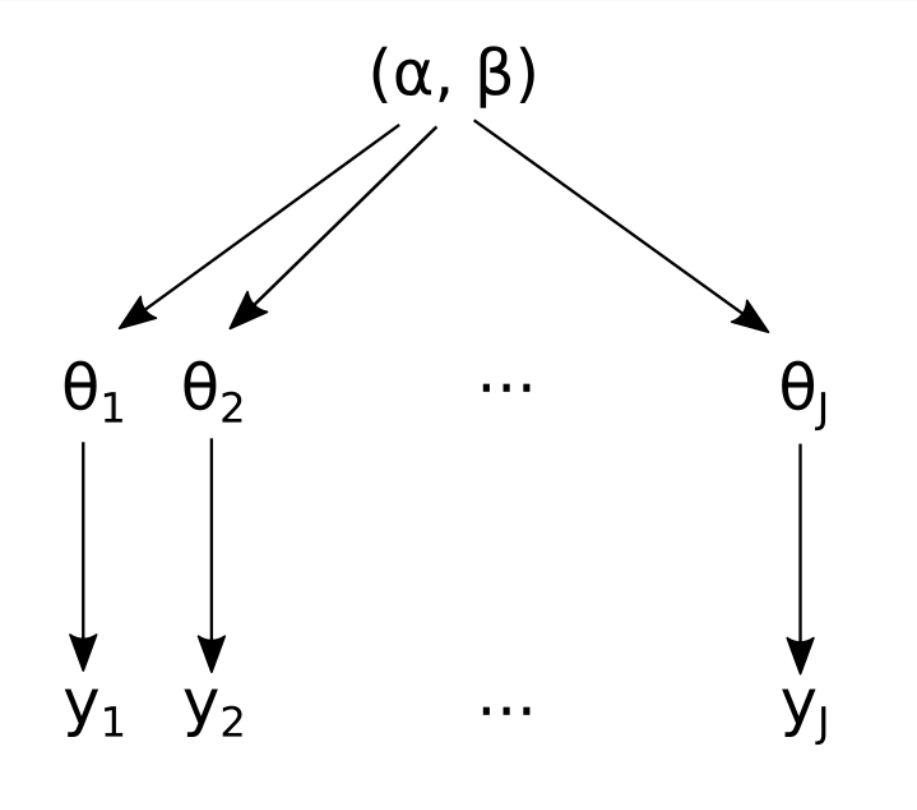
Independence of the θ_js

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 separated 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"

Ignorance implies exchangeability

These exemplify a broad practical idea: ignorance implies exchangeability.

- The less we know about a problem, the stronger a claim of exchangeability
- Example: a die with 6 sides
  - Initially all sides are exchangeable
  - Careful examination of the die might reveal imperfections, leading us to distinguish sides from one another
- If we don't know whether the streets have bike lanes, then they're exchangeable
- If we know that $y_1 0$, the 10th sampled street, is University Blvd., then it shouldn't be treated as exchangeable with the others (we know geographic factors affecting bicycle traffic)

Hierarchical normal model

Example: 8 schools

Example: SAT coaching effectiveness

- SAT design intent: short term coaching should not improve outcomes significantly
- nonetheless, schools implement coaching programs
- examine effectiveness of coaching programs

Experiment:

- All students pre-tested with PSAT
- Some students coached
- Coaching effects y_i estimated with linear regression
- Data is at the school level, not individual

Example: 8 schools

Data:

School	Effect	SE
A	28	15
B	8	10
C	-3	16
D	7	11
E	-1	9
F	1	11
G	18	10
H	12	18

The model

Normals at all levels:

$$
\begin{aligned}
y_j &\sim \text{Normal}(\theta_j, SE_j) \\
\theta_j &\sim \text{Normal}(\mu, \tau) \\
\mu &\sim \text{Normal}(\mu_0, \sigma_0) \\
\tau &\sim \text{HalfCauchy}(5)
\end{aligned}
$$

Notice: take SE to be known, only interested in estimating θ_j.

Draw the model

Computation and computational difficulties

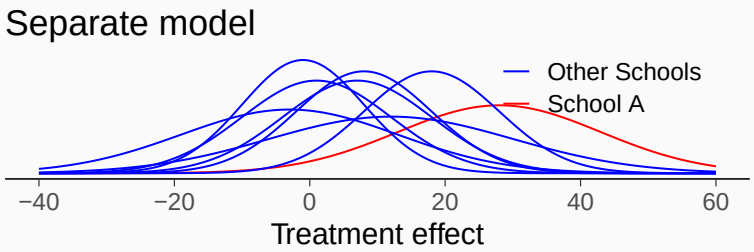
This model is easy to conceptualize, and structurally similar to the bike lane model.

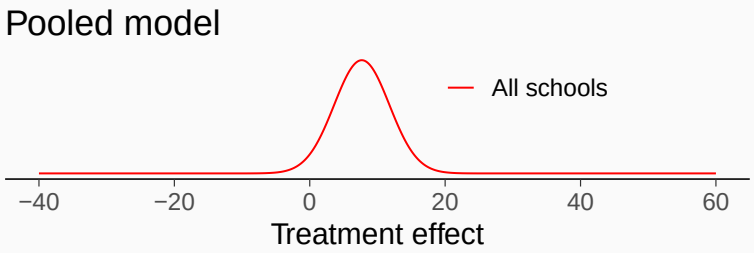
But:

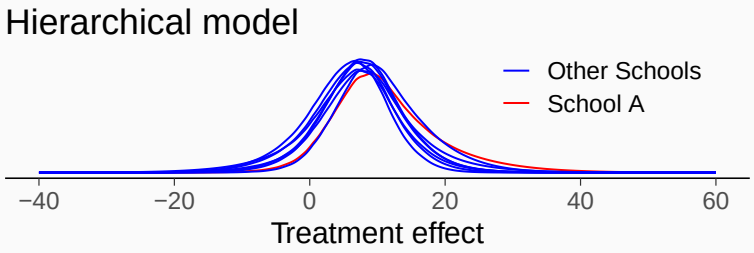
- The hierarchical normal model has some computational challenges
- Difficult for MCMC samplers to explore without re-parameterization

Let's take a look in PyMC3...

Results





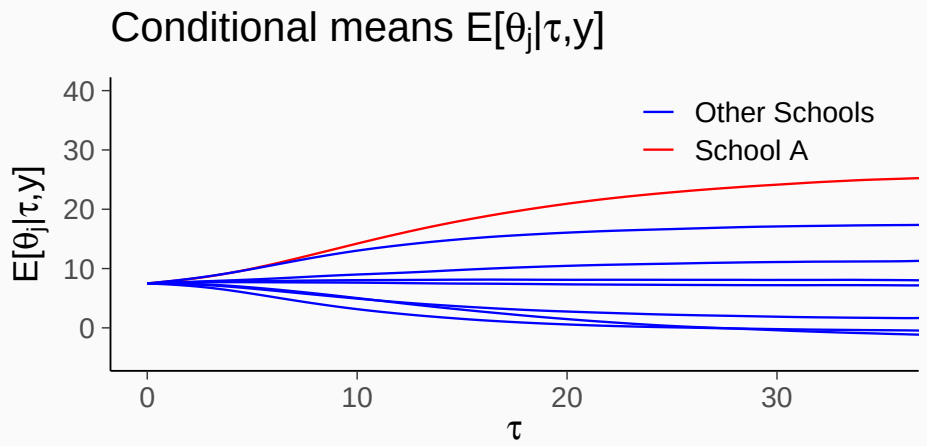


(graphics courtesy Aki Vehtari)

Hierarchical model as a compromise

Remember the (hyper)parameter τ

If we condition on τ:



Hierarchical model is "partial pooling" – compromise between total pooling and separate effects

Amount of pooling controlled by τ; hierarchical model learns this from the data.

Summary

Next week:

- MCMC - what is it?
- How do modern MCMC methods work?
- Diagnosing sampling problems