Intro to Hierarchical Models

ISTA 410 / INFO 510: Bayesian Modeling and Inference

U. of Arizona School of Information

October 11, 2021

Outline

Last week:

- Information criteria: AIC, WAIC
- Approximate leave-one-out cross-validation: PSIS

This week: hierarchical/multilevel models

Intro to hierarchical models

How much traffic is bicycle traffic?

In BDA3 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:

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

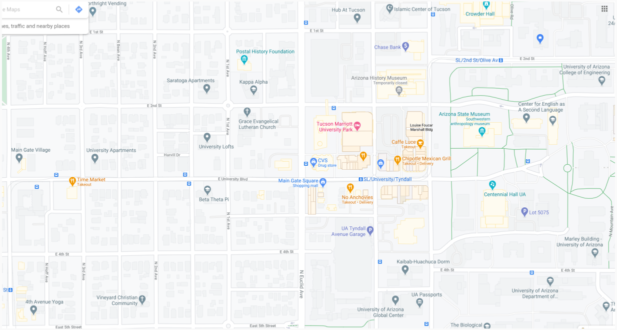
for fixed α_0, β_0.

- 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:

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

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

Posterior distributions

- Both models use conjugacy, so we can write down the posteriors
- Pooled model:

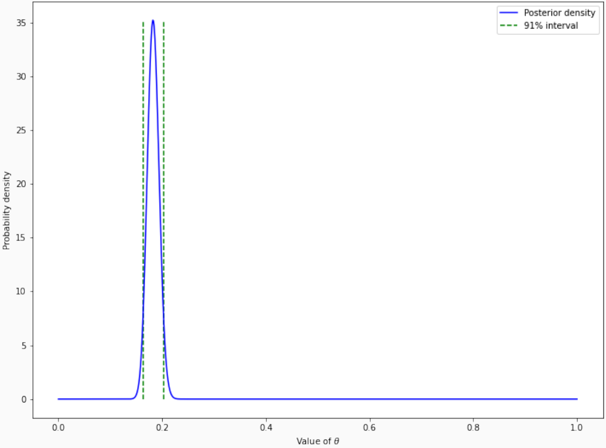
$$\theta | y \sim \mathrm{Beta}(1 + n_{\mathrm{bikes}}, 3 + n_{\mathrm{others}})$$

- Separated model:

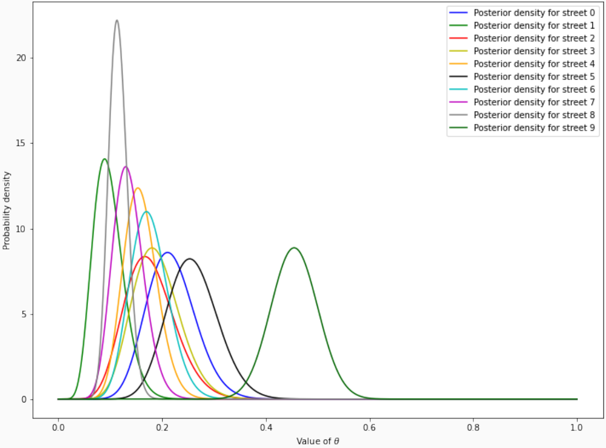
$$\theta_j | y \sim \mathrm{Beta}(1 + n_{\mathrm{bikes},j}, 3 + n_{\mathrm{others},j})$$

	bicycles	others
0	16	58
1	9	90
2	10	48
3	13	57
4	19	103
5	20	57
6	18	86
7	17	112
8	35	273
9	55	64

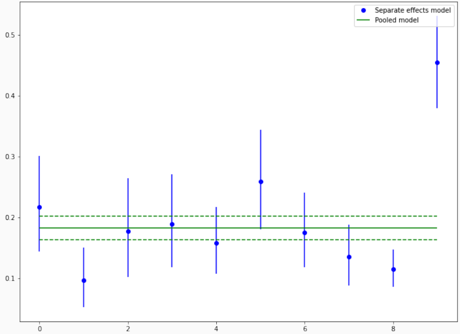
Examine the results



Examine the results



Examine the results



- All error bars are 91% interval
- Is it a conflict that some streets fall outside the pooled interval?

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!

Predicting the next street

What if we wanted to predict the proportion of bike traffic we would see if we went out and observed street 8 again?

- Use the separated model's estimate for street 8
- Pooled model probably overestimates

What if we wanted to predict the proportion of bike traffic we would see if we observed a new street?

- Any individual street from the separated model is unlikely to be a good estimate
- The pooled model has far too little uncertainty

Two scales of variation

We can think of these as modeling two different scales of variation:

- The pooled model treats all streets as equivalent; we're estimating a quantity at a city scale
  - How many people cycle in *this city*?
- The separate model treats all streets as independent entities – really like 10 different models
  - How many people cycle on *this street*?

Two scales of variation

In order to predict the next street, we should understand both scales of variation:

- The pooled model has not learned anything about how different traffic patterns are on different streets
- The separated model has learned nothing about what a typical street looks like
  - When predicting a new street, the separated model just goes back to the prior
- Neither of these is a complete picture

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.

The hierarchical model

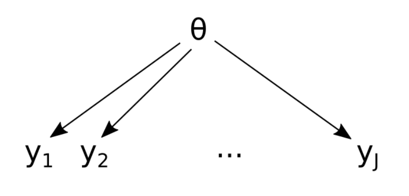
With a Bayesian approach, we can find a compromise.

- We have a θ for each street (like the separated model)
- 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
    - The hyperparameters represent population/city-scale variation (like what is estimated by the pooled model)

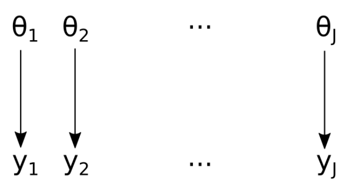
(note: slightly different sense of the term *hyperparameter* from its common use in ML)

Examining this graphically

Pooled model:

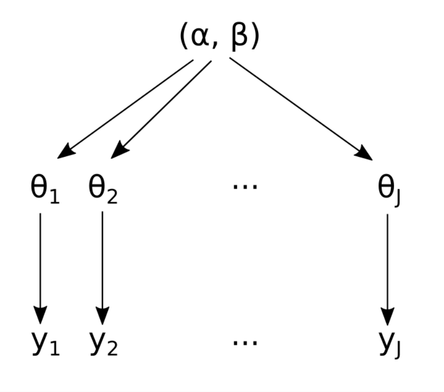


Separate model:



Examining this graphically

Hierarchical model combines the features of these two:



Setting up the model

This is conceptually only a slight difference from the separated model:

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

We just need to set up a prior distribution for α, β

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 always 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:

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

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)

Setting up the model (in PyMC3)

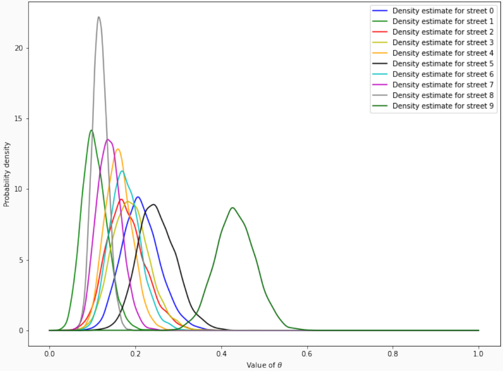
In PyMC3:

```
with pm.Model() as hierarchical_model:
    # Hyperpriors
    mu = pm.Beta('mu', 1, 3)
    eta = pm.HalfCauchy('eta', 1)

    alpha = eta * mu
    beta = eta * (1 - mu)

    # Distributions for theta
    # shape = 10 makes a vector of 10 parameters
    theta = pm.Beta('theta', alpha=alpha, beta=beta, shape = 10)

    # Likelihood
    y_obs = pm.Binomial('y_obs', p = theta, observed = df.bicycles, n=df.total)

    # Inference
    trace = pm.sample()
```

Comparison

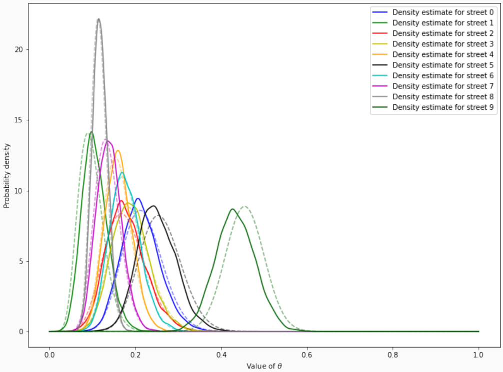
What is the difference in the results?

Comparing posterior densities:



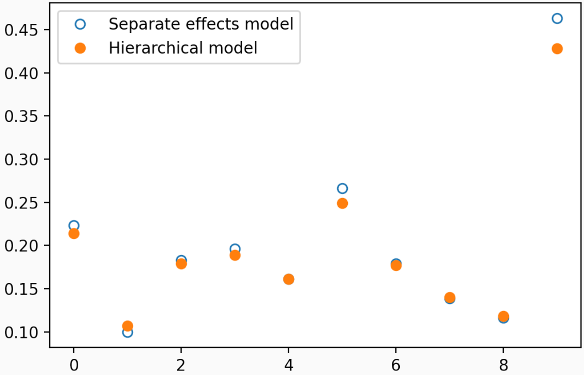
What is the difference in the results?

Comparing posterior densities:



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.

Predicting the next street

Going back to our motivating question:

- What prediction could we make for the rate of bicycle traffic on a newly observed street?
  - Make this concrete: if we go to a new street and observe 100 vehicles, what is a 50% interval for the number of bicycles?
- Multi-level posterior prediction
  - Our new street has a rate θ_{11} drawn from $\mathrm{Beta}(\mu, \eta)$
  - Draw values from the posterior distribution of μ, η; use them to sample a new θ_{11}
  - Then, the number of bicycles is drawn from $\mathrm{Binomial}(100, \theta_{11})$

Uncertainty at multiple scales

This prediction process incorporates three random draws, for three scales of uncertainty:

1. We don't know the values of μ, η that describe the distribution of individual street properties
2. Conditional on μ, η, have uncertainty about the new θ_{11}
3. Conditional on θ_{11} there is uncertainty about how many bikes will pass during our observation

If we were predicting an observation on street 8:

- Still have parts 2 and 3 above, but μ, η no longer needed

Independence and exchangeability

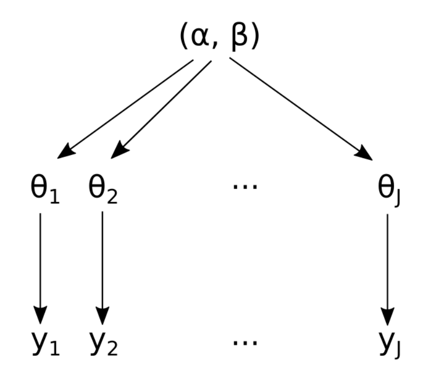
Independence of the θ_js

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:

$$\begin{aligned} w &= \frac{\alpha}{\alpha + \beta} \\ z &= (\alpha + \beta)^{-1/2} \\ p(w, z) &\propto 1 \end{aligned}$$

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)