Exchangeability and more hierarchical models

ISTA 410 / INFO 510: Bayesian Modeling and Inference

U. of Arizona School of Information

October 13, 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

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

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

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:

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

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

Call this the separate-effects model.

Setting up the model

A compromise: hierarchical model

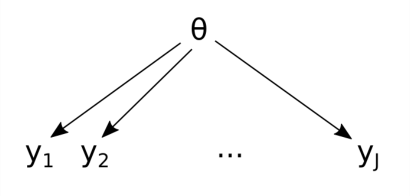
$$
\begin{aligned}
y_j &\sim \text{Binomial}(\theta_j, n_j) \\
\theta_j &\sim \text{Beta}(\alpha, \beta) \\
\mu &:= \frac{\alpha}{\alpha + \beta} \\
\eta &:= \alpha + \beta \\
p(\mu) &\sim \text{Beta}(1, 3) \\
p(\eta) &\sim \text{HalfCauchy}(1)
\end{aligned}
$$

Note: BDA3 uses

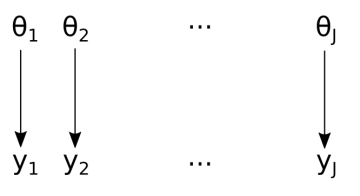
$$p(\alpha, \beta) \propto (\alpha + \beta)^{-5/2}$$

Examining this graphically

Pooled model:

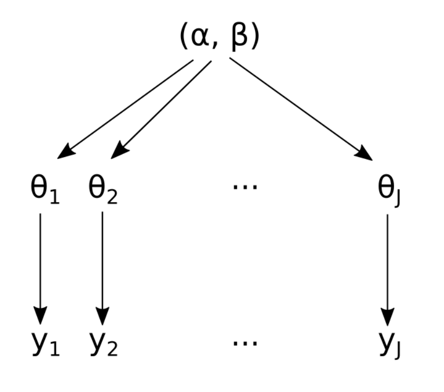


Separate model:



Examining this graphically

Hierarchical model combines the features of these two:



Setting up the model (in PyMC3)

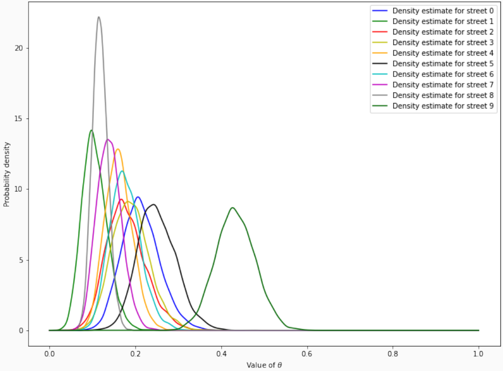
In PyMC3:

```
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

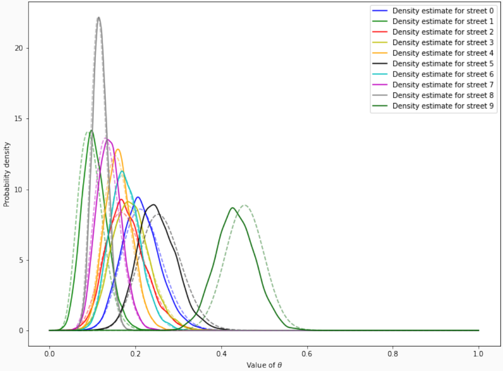
What is the difference in the results?

Comparing posterior densities:



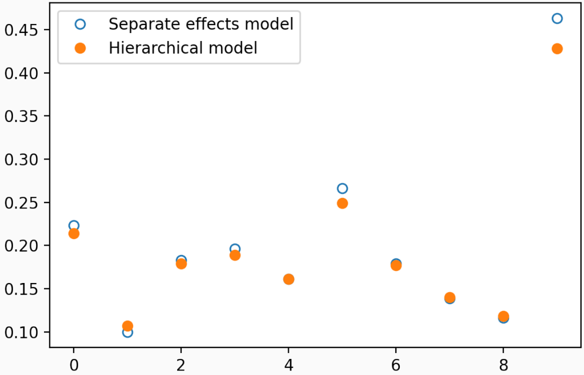
What is the difference in the results?

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

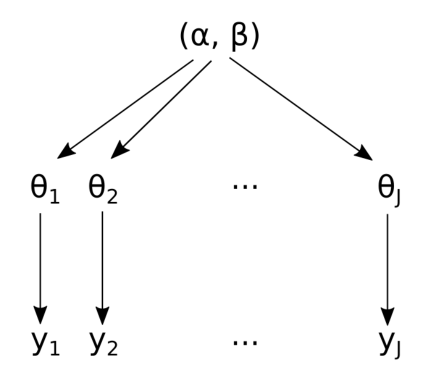
Independence of the θ_js

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 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

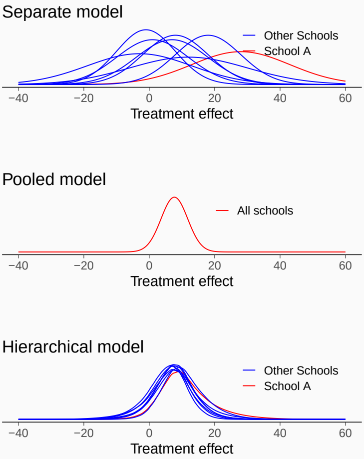
Normals at all levels:

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

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

Draw the model

Results

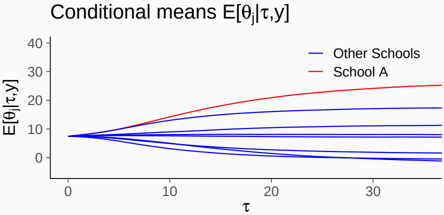


(graphics courtesy Aki Vehtari)

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.

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...

What is a divergence?

The core sampling step of HMC is a physics simulation:

- The Hamiltonian $H(q, p)$ represents the total energy of the system
- Hamiltonian systems conserve energy

A "divergence" in HMC is when the energy at the start of the simulation doesn't match the energy at the end. It means something went wrong with the physics.

Divergence: causes and effects

What causes a divergence: numerical problems.

- The physics simulation uses discrete time steps to model a continuous process
- If our time steps are too large, our simulation is too "coarse"
- Time steps need to be smaller when the potential energy (i.e. log posterior) has high curvature (2nd derivative) meaning that the simulation loses accuracy.

How to fix divergences

When we get divergences, the sampler has some suggestions for us:

```
There were 282 divergences after tuning.
Increase `target_accept` or reparameterize.
```

- increase `target_accept`
- reparameterize

Simple approach: change target acceptance

The simplest way to deal with divergences: decrease the step size in the physics simulation. This is what `target_accept` controls.

- Metropolis step corrects for numerical error – finer resolution means more proposals accepted
- This decreases the efficiency of the sampling, because it requires more physics steps per sample.
- Default for `target_accept` is 0.8; try setting to, e.g., 0.9, 0.95

This works for random "false positive" divergences, but sometimes won't help. Divergences that won't go away are a serious problem.

The folk theorem

The "folk theorem of statistical computing:"

[Gelman] When you have computation problems, often there's a problem with your model.

- Check simple things first. Things I have done that have led to computational problems:
  - Forgetting to set the shape of a parameter
  - Forgetting to link up two parameters
- Set reasonable weakly regularizing priors

Excessive curvature in the 8 schools problem

Where is the problem in the 8 schools model?

Let's first look at some plots.

Excessive curvature in the 8 schools problem

Where is the problem in the 8 schools model?

Let's first look at some plots.

Recall the model:

$$\begin{aligned}
y_j &\sim \mathrm{Normal}(\theta_j, \sigma_j) \\
\theta_j &\sim \mathrm{Normal}(\mu, \tau) \\
\mu &\sim \mathrm{Normal}(0, 5) \\
\tau &\sim \mathrm{HalfCauchy}(5)
\end{aligned}$$

The source of the problem

Remember the physics "surface" is the log posterior. So let's examine that.

The source of the problem

Remember the physics "surface" is the log posterior. So let's examine that.

$$
\begin{aligned}
p(\mu, \tau, \theta | y) &\propto p(\mu, \tau, \theta) p(y | \mu, \tau, \theta) \\
&= p(\mu, \tau) p(\theta | \mu, \tau) p(y | \theta)
\end{aligned}
$$

Taking logs,

$$
\begin{aligned}
logp(\mu, \tau, \theta | y) &= \mathrm{const} + \log p(\mu, \tau) \\
&- \frac{1}{2} \sum_{j=1}^{J} \left(\frac{\theta_j - \mu}{\tau} \right)^2 \\
&- \frac{1}{2} \sum_{j=1}^{J} \left(\frac{y_j - \theta_j}{\sigma_j} \right)^2
\end{aligned}
$$

Reparameterizing the 8 schools

We can reparameterize to a "non-centered" parameterization.

- If $x \sim N(0, 1)$, then $\sigma x \sim N(0, \sigma)$
- So, the following is equivalent to our original model:

$$\begin{aligned}
y_j &\sim \mathrm{Normal}(\theta_j, \sigma_j) \\
\eta_j &\sim \mathrm{Normal}(0, 1) \\
\theta_j &= \mu + \tau \eta_j \\
\mu &\sim \mathrm{Normal}(0, 5) \\
\tau &\sim \mathrm{HalfCauchy}(5)
\end{aligned}$$

We add a latent variable and move all the variance of θ_j into it; τ no longer appears in the denominator in the log posterior.

Diagnostic statistics

Some other diagnostic statistics that you can use:

- $\hat{R}$ (a.k.a. the Gelman-Rubin statistic), measures the ratio of the estimated variance of the parameter, pooling several chains, to the variance within a chain. Should be nearly 1 if all chains have converged to the same distribution. You'll get warnings if $\hat{R}$ is too high for any parameter. (See BDA sec. 11.4)
- Effective sample size: estimate of the equivalent number of samples *without* autocorrelation. You'll get warnings if this is really low. (See BDA sec. 11.5)

Both calculated by `az.summary`; let's take a look.

Summary

Further reading for practical advice on using MCMC methods:

- Divergences in the 8 schools model. https://colcarroll.github.io/pymc3/notebooks/Diagnosing_biased_Inference_with_Divergences.html
- General notes on using HMC in PyMC3. Lots of good small tips and tricks here. https://eigenfoo.xyz/_posts/2018-06-19-bayesian-modelling-cookbook/
- Notes on choosing priors. Written by Stan users, but most of it is really language-agnostic (because it's more about the modeling side). https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations

Summary

Hierarchical models:

- Model variation on multiple scales
- Allow sharing of information for estimates of exchangeable groups

Next week:

- Hierarchical linear models and GLMs