Exchangeability and more hierarchical models

ISTA 410 / INFO 510: Bayesian Modeling and Inference

U. of Arizona School of Information

September 23, 2020

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.

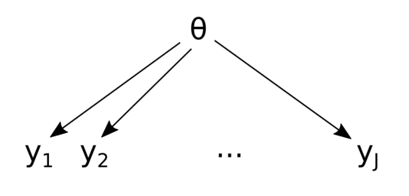
Setting up the model

A compromise: hierarchical model

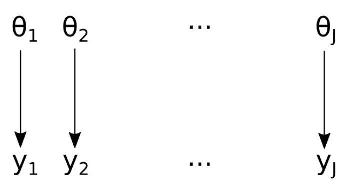
$$
\begin{aligned}
y_j &\sim \mathrm{Binomial}(\theta_j, n_j) \\
\theta_j &\sim \mathrm{Beta}(\alpha, \beta) \\
p(\alpha, \beta) &\propto (\alpha + \beta)^{-5/2}
\end{aligned}
$$

Examining this graphically

Pooled model:

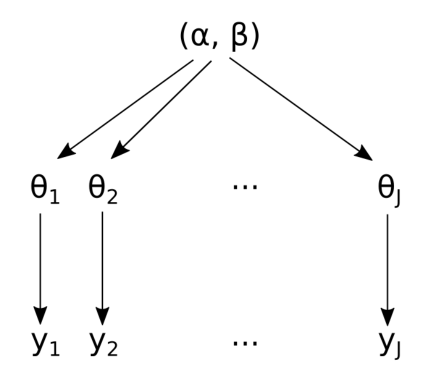


Separate model:



Examining this graphically

Hierarchical model combines the features of these two:



Probabilistic description

Factoring the joint distribution

The graphical model implies the following factorization of the joint probability distribution of all variables:

$$
\begin{aligned}
p(\alpha, \beta, \theta, y) &= p(\alpha, \beta) p(\theta | \alpha, \beta) p(y | \theta, \alpha, \beta) \\
&= p(\alpha, \beta) p(\theta | \alpha, \beta) p(y | \theta)
\end{aligned}
$$

Posteriors

The joint posterior for all the parameters is then written down as

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

So θ mediates the information flow in both directions: y depends on α, β only through its effect on θ; and, when y is observed, $p(\alpha, \beta)$ is updated only through θ's update.

Marginal posterior of hyperparameters

It is useful, however, to find the marginal $p(\alpha, \beta|y)$ describing the update of α, β due to y. Two approaches:

- Direct integration:

$$p(\alpha, \beta|y) = \int p(\alpha, \beta, \theta|y) d\theta$$

- Algebra (when it works):

$$p(\alpha, \beta|y) = \frac{p(\alpha, \beta, \theta|y)}{p(\theta|\alpha, \beta, y)}$$

The second approach is what is used in the bike traffic example (or rat tumor example in book section 5.3)

Exchangeability

Exchangeability

Exchangeability is the justification for applying a joint prior to the parameters in our model.

Formally, θ_i are exchangeable if the joint distribution is invariant under permutations of the index, e.g.,

$$p(\theta_1, \theta_2, \theta_3) = p(\theta_2, \theta_1, \theta_3)$$

Exchangeability is why we can consider the θ parameters as *a priori* the same even though we expect them to ultimately be different

Exchangeability in the bike traffic example

In the bike traffic example:

- Individual observations (vehicles) on each street are exchangeable, but observations
- Streets are exchangeable

Exchangeability in the bike traffic example

In the bike traffic example:

- Individual observations (vehicles) on each street are exchangeable, but observations
- Streets are exchangeable
- Exchangeability does not imply that the streets could not be different. We know it is plausible that some streets are more popular with bicyclists or less popular with cars; but we don't know which streets are which *a priori*

Exchangeability in the bike traffic example

In the bike traffic example:

- Individual observations (vehicles) on each street are exchangeable, but observations
- Streets are exchangeable
- Exchangeability does not imply that the streets could not be different. We know it is plausible that some streets are more popular with bicyclists or less popular with cars; but we don't know which streets are which *a priori*
- If we expand the model to include both streets with and without bike lanes:
  - streets with bike lanes are exchangeable
  - streets without bike lanes are exchangeable
  - all streets are not exchangeable

Another example of exchangeability

From the book:

- Data: y_i = divorce rates per 1000 in 8 US states, but I don't tell you which state
- A priori, y_i are exchangeable; it does not matter which state is $y_1, y_2, \ldots$
- Observe the first seven: 5.6, 6.6, 7.8, 5.6, 7.0, 7.2, 5.4; y_8 still unknown.
  - Still exchangeable; your model wouldn't change if the missing observation were y_1 instead of y_8; that is,
  - $p(y_8|y_1, y_2, \ldots) = p(y_1|y_2, y_3, \ldots)$
  - You'd predict $p(y_8|y_1, y_2, \ldots)$ is centered around 6.5, mostly fall between 5 and 8

Another example of exchangeability

How much more information do we need before the states are not exchangeable

- What if you know that the states are in the Mountain west: AZ, CO, ID, MT, NV, NM, UT, WY?
- Still exchangeable, but probably change the priors: expect a couple of outliers (UT, NV)
- Now we make the observations: 5.6, 6.6, 7.8, 5.6, 7.0, 7.2, 5.4

Another example of exchangeability

How much more information do we need before the states are not exchangeable

- What if you know that the states are in the Mountain west: AZ, CO, ID, MT, NV, NM, UT, WY?
- Still exchangeable, but probably change the priors: expect a couple of outliers (UT, NV)
- Now we make the observations: 5.6, 6.6, 7.8, 5.6, 7.0, 7.2, 5.4
- We suspect the last state is either NV or UT, but we can't say which; prior still exchangeable

Another example of exchangeability

How much more information do we need before the states are not exchangeable

- What if you know that the states are in the Mountain west: AZ, CO, ID, MT, NV, NM, UT, WY?
- Still exchangeable, but probably change the priors: expect a couple of outliers (UT, NV)
- Now we make the observations: 5.6, 6.6, 7.8, 5.6, 7.0, 7.2, 5.4
- We suspect the last state is either NV or UT, but we can't say which; prior still exchangeable
- Finally, say we know y_8 is Nevada. Exchangeable?

Another example of exchangeability

How much more information do we need before the states are not exchangeable

- What if you know that the states are in the Mountain west: AZ, CO, ID, MT, NV, NM, UT, WY?
- Still exchangeable, but probably change the priors: expect a couple of outliers (UT, NV)
- Now we make the observations: 5.6, 6.6, 7.8, 5.6, 7.0, 7.2, 5.4
- We suspect the last state is either NV or UT, but we can't say which; prior still exchangeable
- Finally, say we know y_8 is Nevada. Exchangeable?
- No; even before seeing the 7 values we cannot assign an exchangeable prior, and after observation we should place most posterior probability for y_8 above (e.g.) 8

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

Exchangeability vs. independence

- Imagine a die with 6 sides. $\theta_i = p(\text{roll } i)$. Are θ_i exchangeable? Independent?

Exchangeability vs. independence

- Imagine a die with 6 sides. $\theta_i = p(\text{roll } i)$. Are θ_i exchangeable? Independent?
- Exchangeable: yes – before rolling the die, we don't necessarily think all θ_i are the same, but we can't distinguish between them

Exchangeability vs. independence

- Imagine a die with 6 sides. $\theta_i = p(\text{roll } i)$. Are θ_i exchangeable? Independent?
- Exchangeable: yes – before rolling the die, we don't necessarily think all θ_i are the same, but we can't distinguish between them
- Independent: no – any 5 determine the 6th because of the constraint $\sum \theta_i = 1$.

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

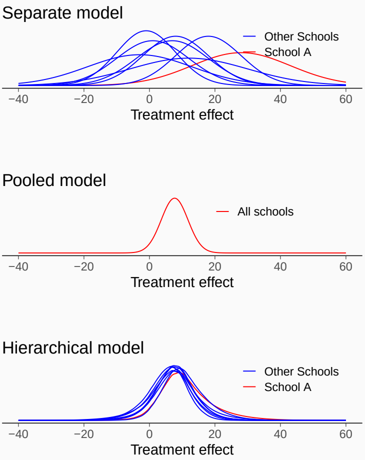
Normals at all levels:

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

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

Draw the model

Results

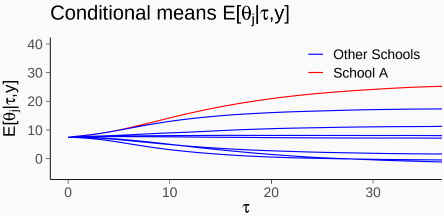


(graphics courtesy Aki Vehtari)

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.

Summary

Next week:

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