Hierarchical models (cont.)

Dr. Jarad Niemi

STAT 544 - Iowa State University

February 14, 2018

Outline

- Theoretical justification for hierarchical models
  - Exchangeability
  - de Finetti's theorem
  - Application to hierarchical models
- Normal hierarchical model
  - Posterior
  - Simulation study
  - Shrinkage

Exchangeability

Definition

The set $Y_1, Y_2, \ldots, Y_n$ is exchangeable if the joint probability $p(y_1, \ldots, y_n)$ is invariant to permutation of the indices. That is, for any permutation π,

$$p(y_1, \ldots, y_n) = p(y_{\pi_1}, \ldots, y_{\pi_n}).$$

An exchangeable but not iid example:

- Consider an urn with one red ball and one blue ball with probability $1/2$ of drawing either.
- Draw without replacement from the urn.
- Let $Y_i = 1$ if the ith ball is red and otherwise $Y_i = 0$.
- Since $1/2 = P(Y_1 = 1, Y_2 = 0) = P(Y_1 = 0, Y_2 = 1) = 1/2$, Y_1 and Y_2 are exchangeable.
- But $0 = P(Y_2 = 1 | Y_1 = 1) \neq P(Y_2 = 1) = 1/2$ and thus Y_1 and Y_2 are not independent.

Exchangeability

Theorem

All independent and identically distributed random variables are exchangeable.

Proof.

Let $y_i \stackrel{iid}{\sim} p(y)$, then

$$p(y_1, \ldots, y_n) = \prod_{i=1}^{n} p(y_i) = \prod_{i=1}^{n} p(y_{\pi_i}) = p(y_{\pi_1}, \ldots, y_{\pi_n})$$

□

Definition

The sequence $Y_1, Y_2, \ldots$ is infinitely exchangeable if, for any n, $Y_1, Y_2, \ldots, Y_n$ are exchangeable.

de Finetti's theorem

Theorem

A sequence of random variables $(y_1, y_2, \ldots)$ is infinitely exchangeable iff, for all n,

$$p(y_1, y_2, \ldots, y_n) = \int \prod_{i=1}^{n} p(y_i|\theta) P(d\theta),$$

for some measure P on θ.

If the distribution on θ has a density, we can replace $P(d\theta)$ with $p(\theta)d\theta$.

This means that there must exist

- a parameter θ,
- a likelihood $p(y|\theta)$ such that $y_i \stackrel{ind}{\sim} p(y|\theta)$, and
- a distribution P on θ.

Application to hierarchical models

Assume $(y_1, y_2, \ldots)$ are infinitely exchangeable, then by de Finetti's theorem for the $(y_1, \ldots, y_n)$ that you actually observed, there exists

- a parameter θ,
- a distribution $p(y|\theta)$ such that $y_i \stackrel{ind}{\sim} p(y|\theta)$, and
- a distribution P on θ.

Assume $\theta = (\theta_1, \theta_2, \ldots)$ with θ_i infinitely exchangeable. By de Finetti's theorem for $(\theta_1, \ldots, \theta_n)$, there exists

- a parameter ϕ,
- a distribution $p(\theta|\phi)$ such that $\theta_i \stackrel{ind}{\sim} p(\theta|\phi)$, and
- a distribution P on ϕ.

Assume $\phi = \phi$ with $\phi \sim p(\phi)$.

Exchangeability with covariates

Suppose we observe y_i observations and x_i covariates for each unit i. Now we assume $(y_1, y_2, \ldots)$ are infinitely exchangeable given x_i, then by de Finetti's theorem for the $(y_1, \ldots, y_n)$, there exists

- a parameter θ,
- a distribution $p(y|\theta, x)$ such that $y_i \stackrel{ind}{\sim} p(y|\theta, x_i)$, and
- a distribution P on θ given x.

Assume $\theta = (\theta_1, \theta_2, \ldots)$ with θ_i infinitely exchangeable given x. By de Finetti's theorem for $(\theta_1, \ldots, \theta_n)$, there exists

- a parameter ϕ,
- a distribution $p(\theta|\phi, x)$ such that $\theta_i \stackrel{ind}{\sim} p(\theta|\phi, x_i)$, and
- a distribution P on ϕ given x.

Assume $\phi = \phi$ with $\phi \sim p(\phi|x)$.

Summary

Hierarchical model:

$$y_i \stackrel{ind}{\sim} p(y|\theta_i), \qquad \theta_i \stackrel{ind}{\sim} p(\theta|\phi), \qquad \phi \sim p(\phi)$$

Hierarchical linear model:

$$y_i \stackrel{ind}{\sim} p(y|\theta_i, x_i), \qquad \theta_i \stackrel{ind}{\sim} p(\theta|\phi, x_i), \qquad \phi \sim p(\phi|x)$$

Although hierarchical models are typically written using the conditional independence notation above, the assumptions underlying the model are exchangeability and functional forms for the priors.

Normal hierarchical models

Suppose we have the following model

$$\begin{aligned} y_{ij} &\stackrel{ind}{\sim} N(\theta_i, \sigma^2) \\ \theta_i &\stackrel{iid}{\sim} N(\mu, \tau^2) \end{aligned}$$

with $j = 1, \ldots, n_i$, $i = 1, \ldots, \mathrm{I}$, and $n = \sum_{i=1}^{J} n_i$. This is a normal hierarchical model.

Make the following assumptions for computational reasons:

- Let $\sigma^2 = s^2$ be known.
- Assume $p(\mu, \tau) \propto p(\mu|\tau)p(\tau) \propto p(\tau)$, i.e. assume an improper uniform prior on μ.

Posterior distribution

The posterior is

$$p(\theta, \mu, \tau | y) \propto p(y|\theta) p(\theta|\mu, \tau) p(\mu|\tau) p(\tau)$$

but the decomposition

$$p(\theta, \mu, \tau | y) = p(\theta|\mu, \tau, y) p(\mu|\tau, y) p(\tau|y)$$

where

$$\begin{array}{ll} p(\theta|\mu, \tau, y) & \propto p(y|\theta)p(\theta|\mu, \tau) \\ p(\mu|\tau, y) & \propto \int p(y|\theta)p(\theta|\mu, \tau) d\theta \, p(\mu|\tau) \\ p(\tau|y) & \propto \int p(y|\theta)p(\theta|\mu, \tau)p(\mu|\tau) d\theta d\mu \, p(\tau) \end{array}$$

will aide computation via

1. $\tau^{(k)} \sim p\left(\tau|y\right)$
2. $\mu^{(k)} \sim p\left(\mu|\tau^{(k)}, y\right)$
3. $\theta_i^{(k)} \sim p\left(\theta|\mu^{(k)}, \tau^{(k)}, y\right)$ for $i = 1, \ldots, \mathrm{I}$.

Posterior distributions

The necessary conditional and marginal posteriors are presented in section 5.4 of BDA. Let

$$\overline{y}_{i\cdot} = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij} \qquad \text{and} \qquad s_i^2 = s^2/n_i$$

Then

$$\begin{array}{rl} p(\tau|y) & \propto p(\tau) V_\mu^{1/2} \prod_{i=1}^{\mathrm{I}} (s_i^2+\tau^2)^{-1/2} \exp\left(-\frac{(\overline{y}_{i\cdot}-\hat{\mu})^2}{2(s_i^2+\tau^2)}\right) \\ \mu|\tau,y & \sim N(\hat{\mu}, V_\mu) \\ \theta_i|\mu,\tau,y & \sim N(\hat{\theta}_i, V_i) \end{array}$$

$$\begin{array}{rlrl} V_\mu^{-1} & = \sum_{j=1}^{J} \frac{1}{s_i^2+\tau^2} & \hat{\mu} & = V_\mu \left(\sum_{i=1}^{\mathrm{I}} \frac{\overline{y}_{\cdot i}}{s_i^2+\tau^2} \right) \\ V_i^{-1} & = \frac{1}{s_i^2} + \frac{1}{\tau^2} & \hat{\theta}_i & = V_i \left(\frac{\overline{y}_{i\cdot}}{s_i^2} + \frac{\mu}{\tau^2} \right) \end{array}$$

Simulation study

Simulation

1. $\theta_i = 0$ for all i
2. $\theta_i = i - (\mathrm{I}/2 + .5)$

Common to both simulations

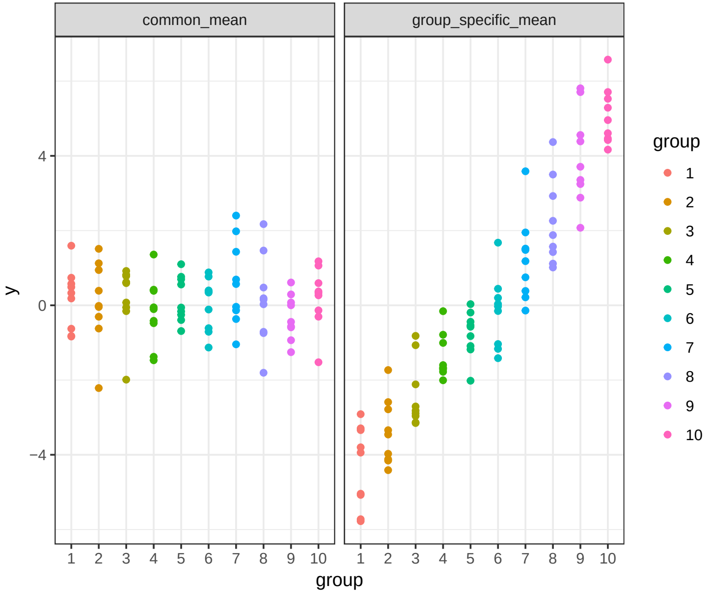
- $\mathrm{I} = 10$
- $n_i = 9$ for all i
- $s = 1$ thus $s_i = 1/3$ for all i

Use $\tau \sim Ca^+(0, 1)$.

Simulation study

```r
J = 10
n_per_group = 9
n = rep(n_per_group,J)
sigma = 1
N = sum(n)
group = rep(1:J, each=n_per_group)

set.seed(1)
df = rbind(data.frame(group = factor(group),
                      simulation = "common_mean",
                      y = rnorm(N                )), # All means are the same
           data.frame(group = factor(group),
                      simulation = "group_specific_mean",
                      y = rnorm(N, group-(J/2+.5)))) # Each group has its own mean
```


common_mean
group_specific_mean
y
4
0
−4
1 2 3 4 5 6 7 8 9 10
1 2 3 4 5 6 7 8 9 10
group
group
1
2
3
4
5
6
7
8
9
10

Summary statistics

	simulation	group	n	mean	sd
1	common_mean	1	9	0.18	0.81
2	common_mean	2	9	0.09	1.11
3	common_mean	3	9	0.18	0.91
4	common_mean	4	9	-0.19	0.89
5	common_mean	5	9	0.17	0.62
6	common_mean	6	9	0.02	0.70
7	common_mean	7	9	0.61	1.14
8	common_mean	8	9	0.14	1.19
9	common_mean	9	9	-0.31	0.60
10	common_mean	10	9	0.20	0.81
11	group_specific_mean	1	9	-4.32	1.10
12	group_specific_mean	2	9	-3.40	0.88
13	group_specific_mean	3	9	-2.41	0.89
14	group_specific_mean	4	9	-1.38	0.60
15	group_specific_mean	5	9	-0.76	0.61
16	group_specific_mean	6	9	-0.16	0.95
17	group_specific_mean	7	9	1.21	1.12
18	group_specific_mean	8	9	2.23	1.15
19	group_specific_mean	9	9	3.97	1.26
20	group_specific_mean	10	9	5.08	0.77

Sampling on a grid

Consider samping from an arbitrary unnormalized density $f(\tau) \propto p(\tau|y)$ using the following approach

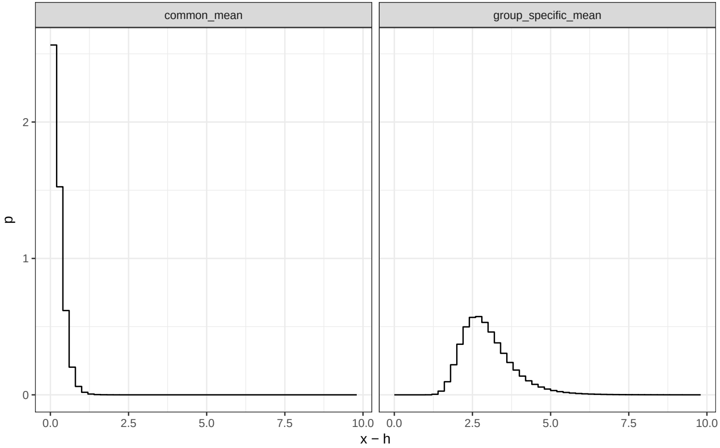
1. Construct a step-function approximation to this density:
   a. Determine an interval $[L, U]$ such that outside this interval $f(\tau)$ is small.
   b. Set an interval half-width h to generate a grid of M points $(x_1, \ldots, x_M)$ in this interval, i.e.

   $$x_1 = L + h \text{ and } x_m = x_{m-1} + 2h \quad \forall\, 1 < m \leq M.$$

   c. Evaluate the density on this grid, i.e. $f(x_m)$.
   d. Normalize interval weights, i.e. $w_m = f(x_m) \Big/ \sum_{i=1}^{M} f(x_i)$

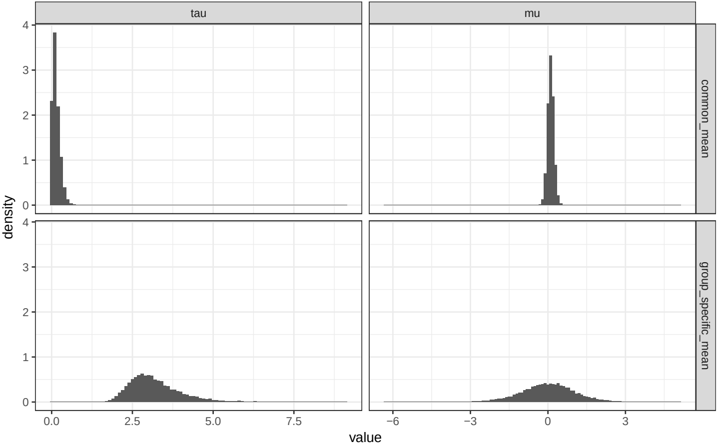
   (to constructed a normalized density, divide each w_m by $2h$.).
2. Sampling from this approximation:
   a. Sample an interval m with probability w_m.
   b. Sample uniformly within this interval, i.e. $\tau \sim \text{Unif}(x_m - h, x_m + h)$.

Approximation to $p(\tau|y)$ when $\tau \sim Ca^+(0,1)$



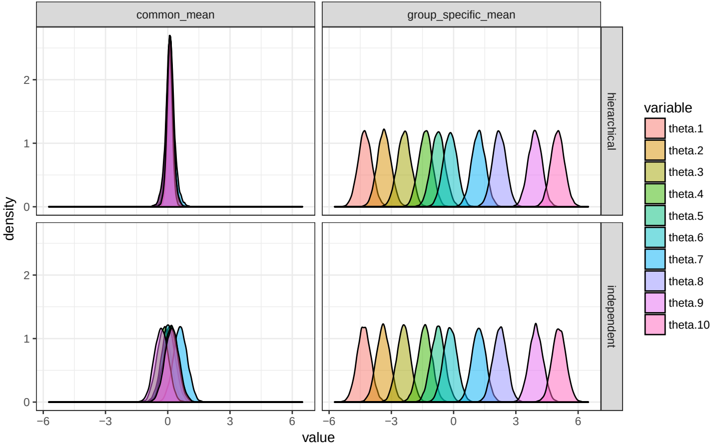
Hyperparameters: group-to-group mean variability

Recall $\theta_i \stackrel{ind}{\sim} N(\mu, \tau^2)$:



Group-specific means

Recall $\theta_i \stackrel{ind}{\sim} N(\mu, \tau^2)$:



Extensions

- Unknown data variance:

$$y_{ij} \sim N(\theta_i, \sigma^2),\, \theta_i \sim N(\mu, \tau^2)$$

or

$$y_{ij} \sim N(\theta_i, \sigma^2),\, \theta_i \sim N(\mu, \sigma^2\tau^2)$$

- Alternative distributions:
  - Heavy-tailed:

$$y_{ij} \sim N(\theta_i, \sigma^2),\, \theta_i \sim t_\nu(\mu, \tau^2)$$

  - Peak at zero:

$$y_{ij} \sim N(\theta_i, \sigma^2),\, \theta_i \sim \text{Laplace}(\mu, \tau^2)$$

  - Point mass at zero:

$$y_{ij} \sim N(\theta_i, \sigma^2),\, \theta_i \sim \pi\delta_0 + (1-\pi)N(\mu, \tau^2)$$

Summary

Hierarchical models

- allow the data to inform us about similarities across groups
- provide data driven shrinkage toward a grand mean
  - lots of shrinkage when means are similar
  - little shrinkage when means are different

Computation used the decomposition

$$p(\theta, \mu, \tau|y) = p(\theta|\mu, \tau, y)p(\mu|\tau, y)p(\tau|y)$$

which allowed for simulation from τ then μ and then θ to obtain samples from the posterior.