Lecture 5: Hierarchical Modeling

Professor Alexander Franks

Announcements

- Reading: Chapter 5 of BDA
    - Ch. 10/11
    - Hw due Sun.

Comparing Multiple Related Groups

- Hierarchy of nested populations
- Models which account for this are called *hiearchical* or *multi-level* models

Some examples:

- Patient outcomes within several different hospitals
- People within counties in the United States (e.g. Asthma mortality example)
- Athlete performance in sports
- Genes within a group of animals

Eight schools example

- A study was performed for the Educational Testing Service (ETS) to evaluate the effects of coaching programs on SAT preparation
- Each of eight different schools used a short-term SAT prep coaching program
- Compute the average SAT score in those who did take the program minus those that did not participate in the program
- We observe the average difference varies by school. What accounts for these differences?

$Y_1, \ldots, Y_8$

- Socioeconomic factors.
- Student baseline
- Hidden versions of treat.
- Sampling variability.

Eight schools example

- Interested in “real” differences due to training
- Want to reduce effect of chance variability
- How do we estimate the effect of the program in each of the schools?

Eight schools example

- Consider two extremes:
  - Estimate the effect of the program in every school independently
    - A separate prior distribution for each school effect
  - Or assume the effect is the same in every school
    - Combine all the data — More data → more power!
  - A compromise between the above 2 options?

Overall estimate of training

$$\begin{cases} Y_j \sim N(\theta_j, \sigma_j^2) \\ \sigma_j^2 \text{ is known.} \\ \sigma_j^2 \propto \dfrac{\sigma^2}{n_j} \end{cases}$$

<u>No pooling</u>: $\hat{\theta}_{MLE,j} = Y_j$

Eight Schools Example

$$y_j \sim N(\theta_j, \sigma_j^2)$$

- y_j is the observed effects of the program in school j
  - Based on a sample of test scores from those in the program and those not in the program
- θ_j are the true *unknown* effects of the program in school j
- Assume variances, σ_j^2, are *known*
  - e.g. determined by the number of students in the sample

Eight Schools Example

```r
J <- 8
y = c(28,  8, -3,  7, -1,  1, 18, 12)
sigma <- c(15, 10, 16, 11,  9, 11, 10, 18)
```

- Assuming the effect of the program on each school is identical.
- What are the chances of seeing a value as large as 28?
- As small as -3?

$$Y_j \sim N(\mu, \sigma_j^2)$$

"complete pooling"

$$\hat{\mu}_{MLE} = \sum_{i=1}^{8} \frac{1/\sigma_i^2}{\sum 1/\sigma_j^2} y_i$$

Eight Schools Example

- Assume the effect of the program on each school is identical, i.e. $\theta_j = \mu$
- Assume a flat prior on μ, what is $p(\mu \mid y_1, \ldots, y_8, \sigma_1, \ldots \sigma_8)$?

$P(\mu) \propto \text{const.}$

$P(\mu \mid y_1, \ldots y_8, \sigma_1, \ldots \sigma_8) \propto L(\mu)$

$\prod^{8} \cancel{\frac{1}{\sqrt{2\pi}\sigma_i}} e^{-\frac{(y_i-\mu)^2}{2\sigma_i^2}} \propto \exp\left[-\sum \frac{1}{2\sigma_i^2}(y_i-\mu)^2\right]$

$$\propto \mathcal{N}\left(\overset{\text{MLE of } \hat{\mu}}{\sum w_i y_i}, \ \frac{1}{\sum 1/\sigma_i^2}\right)$$

$$w_i = \frac{1/\sigma_i^2}{\sum 1/\sigma_j^2}$$

<u>Fisher Weighting</u>.

Eight Schools Example

```r
## Compute the precision frome each school
prec <- 1/sigma^2

## global estimate is a weighted vareage
mu_global <- sum(prec * y / sum(prec))
mu_global
```

```
[1] 7.685617
```

$\hat{\mu}_{MLE}$.

Eight Schools Example

- Assume the effect of the program on each school is identical, i.e. $\theta_j = \mu$
- What are the chances of school 1 having an effect large as 28 (given $\sigma_1 = 15$)?
- Y_3 as small as -3 (given $\sigma_3 = 16$)?

Posterior Prediction Under Complete Pooling

```r
prec <- 1/sigma^2

## global estimate is a weighted average
mu_global <- sum(prec * y / sum(prec))

print(sprintf("mu is %f", mu_global))
```

```
[1] "mu is 7.685617"
```

7.7

```r
1 - pnorm(28, mean=mu_global, sd=sqrt(1/sum(1/sigma^2)+sigma[1]^2))
```

```
[1] 0.09560784
```

```r
pnorm(-3, mean=mu_global, sd=sqrt(1/sum(1/sigma^2)+sigma[3]^2))
```

```
[1] 0.2587447
```

$$P(\tilde{y} \mid \ldots) \sim N(\hat{\mu}_{MLE}, \Sigma 1/\sigma_j^2 + \sigma_1^2)$$

Eight Schools Example

$$y_j \sim N(\theta_j, \sigma_j^2)$$

- θ_j are the true unknown effects of the program in school j
- y_j is the observed effects of the program in school j
  - Based on a sample of test scores from those in the program and those not in the program
  - Number of people in the sample determine the magnitude of σ_j^2

Eight Schools Example

How do we estimate θ_j?

- Assume effects are totally independent: $\hat{\theta}_j^{(MLE)} = y_j$ is the MLE
- Assume effects are identical: $\hat{\theta}_j^{(pool)} = \frac{\sum_i \frac{1}{\sigma_i^2} y_i}{\sum \frac{1}{\sigma_i^2}}$
    - Same effect for all schools: estimate using a weighted average of the observed effects

Eight Schools

```r
theta_j_mle <- y
theta_j_mle
```

```
[1] 28  8 -3  7 -1  1 18 12
```

```r
theta_j_pooled <- rep(sum(1/sigma^2 * y) / sum(1/sigma^2), J)
theta_j_pooled
```

```
[1] 7.685617 7.685617 7.685617 7.685617 7.685617 7.685617 7.685617 7.685617
```

- Compromise: $\hat{\theta}_j^{\text{shrink}} = w\overset{y_j}{\theta_j^{\text{MLE}}} + (1-w)\overset{7.7}{\theta^{pooled}}$

Eight schools example

Add a *shared* normal prior distribution to θ_j

Hierarchical Model

$$\theta_i \sim N(\mu, \tau^2)$$

$$y_j \sim N(\theta_j, \sigma_j^2)$$

Shared Prior. w/ μ (and τ^2)

- The global mean, μ, is also an unknown parameter. What prior should we choose?

also unknown

- τ^2 determines how much weight weight we put on the independent estimate vs the pooled estimate.
- A 9-parameter posterior:
  $p(\mu, \theta_1, \ldots, \theta_8 \mid y_1, \ldots, y_8, \sigma_1, \ldots, \sigma_8, \tau^2)$

Intuition for shrinkage

- $Y_j = \theta_j + \epsilon_j$
  - For simplicity assume $Var(\epsilon_j) = \sigma^2$ for all j
  - θ_j represents true effect in school j (signal)
  - $Var(\theta_j) = \tau^2$ represents how much the true effects vary across schools
    - ϵ_j is sampling variability (noise, chance variation)
- $\hat{\theta}_{MLE} = Y_j$

$$Var(Y_j) = Var(\theta_j + \epsilon_j)$$
$$= \sigma^2 + \tau^2$$

Intuition for shrinkage

- Consequence: the observed outcomes always have higher variance than the signal, i.e. $\mathrm{Var}(Y_j) > \mathrm{Var}(\theta_j)$
- Intuition: reduce the variance by shrinking estimates to a common mean!
- The variance of the shrunken estimates should be close to τ^2

Eight schools example

Questions:

- Is the training program effective in school j?
  - What is $P(\theta_j > 0 \mid y)$?
- On average (over all schools) is the training program effective?
  - What is $P(\mu > 0 \mid y)$?
- Will the training program be effective in a new school?
  - What is $P(\theta_{J+1} > 0 \mid y)$?

Eight schools example

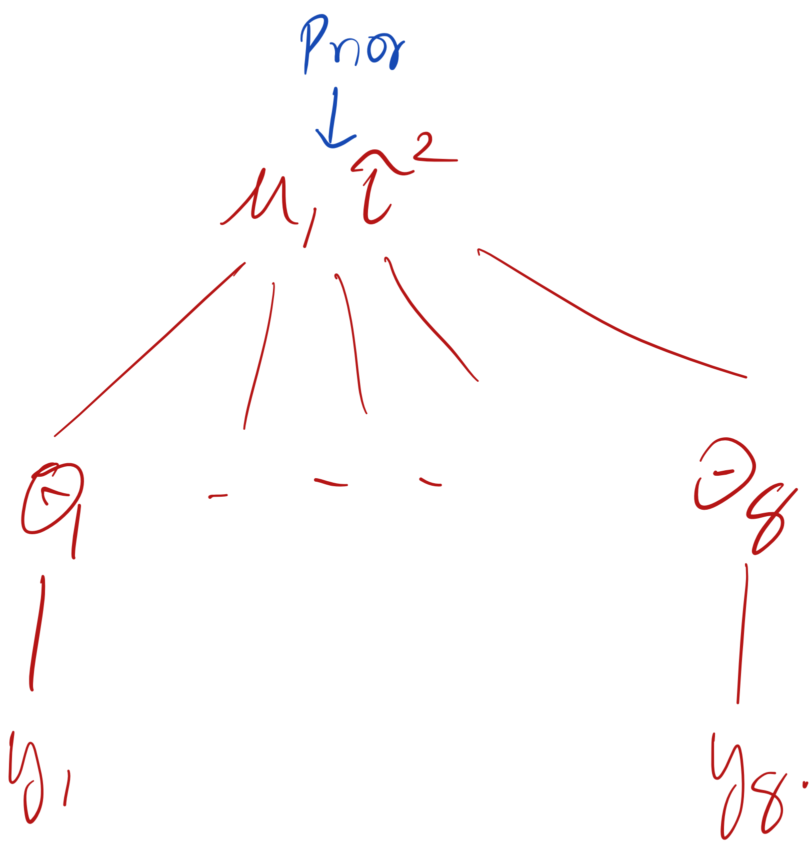
Comments:

- The global average, μ, is a parameter so also has uncertainty
- How dow we determine how much to shrink, e.g. how do we determine τ^2?
- What σ_j^2 were also unknown?

$$y_i \sim N(\theta_i, \sigma_i^2)$$
$$\theta_i \sim N(\mu, \tau^2)$$

If μ & τ^2 are known:

$$p(\theta_i \mid y_1, \ldots y_8, \ldots) \sim N\left(w y_i + (1-w)\mu, \frac{1}{1/\sigma_i^2 + 1/\tau^2}\right)$$



$$w = \frac{1/\sigma_i^2}{1/\sigma_i^2 + 1/\tau^2}$$

Eight schools example

- If τ^2 is large, the prior for θ_j is not very strong
  - If $\tau^2 \to \infty$ equivalent to the no pooling model
- If τ^2 is small, we assume a priori that θ_j are very close
  - if $\tau^2 \to 0$ equivalent to the complete pooling model, $\theta_j = \mu$

$$p(\theta_i \mid \cdots) \sim N\left(w y_i + (1-w)\mu, \frac{1}{1/\sigma_i^2 + 1/\tau^2}\right)$$

$$w = \frac{1/\sigma_i^2}{1/\sigma_i^2 + 1/\tau^2}$$

Inference

- Factorize the density into tractable components

1. ▪ $p(\mu \mid y_1, \ldots, y_8, \tau^2)$ $P(\mu) \propto const.$

2. ▪ $p(\theta_i \mid \mu, y_i, \tau^2)$

- Later: MCMC or other approximate methods

$$P(\mu, \theta_1, \ldots \theta_8 \mid y, \sigma^2, \tau^2) \propto$$

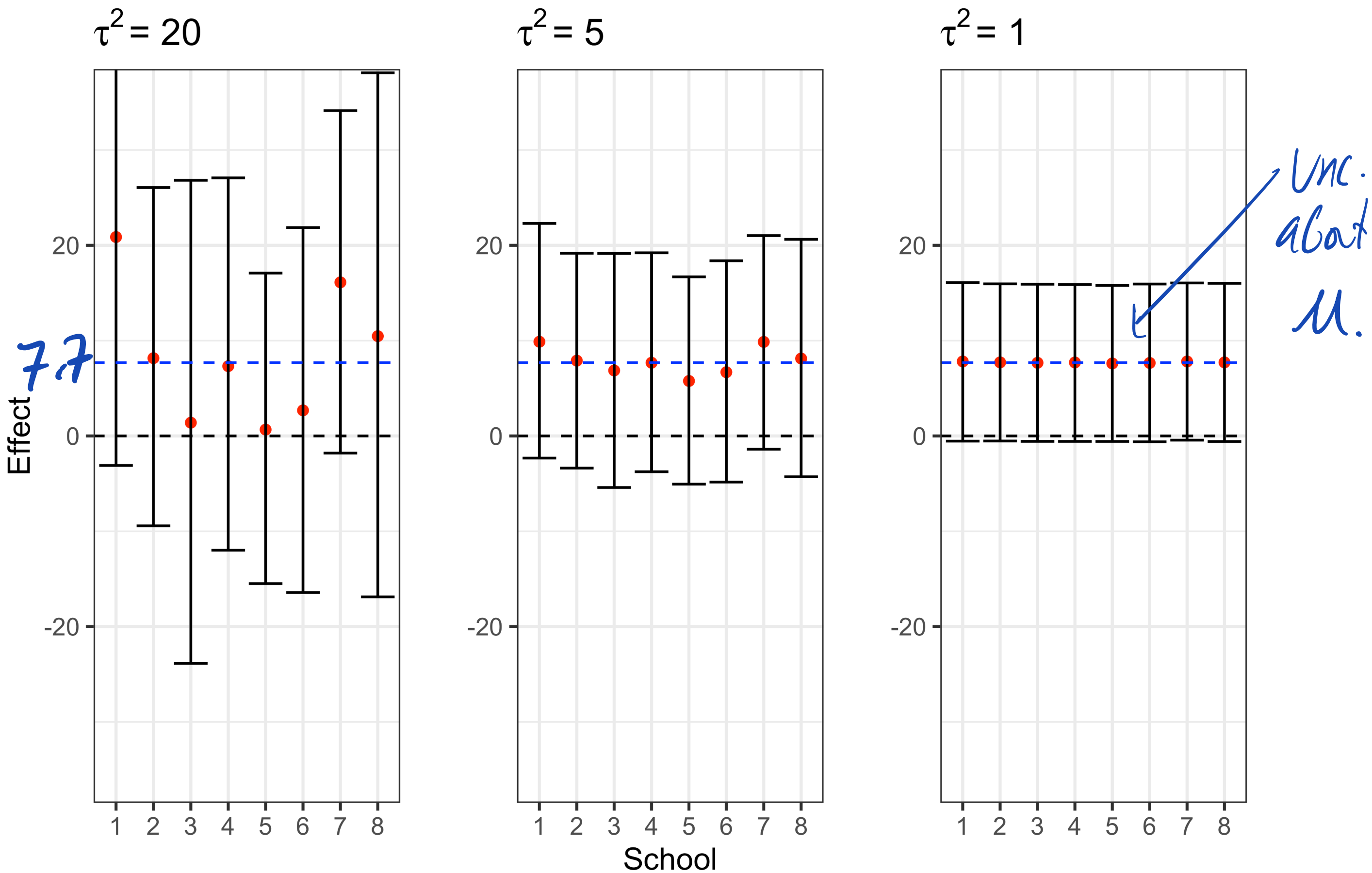
$$P(\mu \mid y, \sigma^2, \tau^2) \underbrace{\prod_{i=1}^{8} P(\theta_i \mid y_i, \sigma_i^2, \mu, \tau^2)}_{P(w y_i + (1-w)\mu, \ \cdot)}$$

$$\downarrow$$

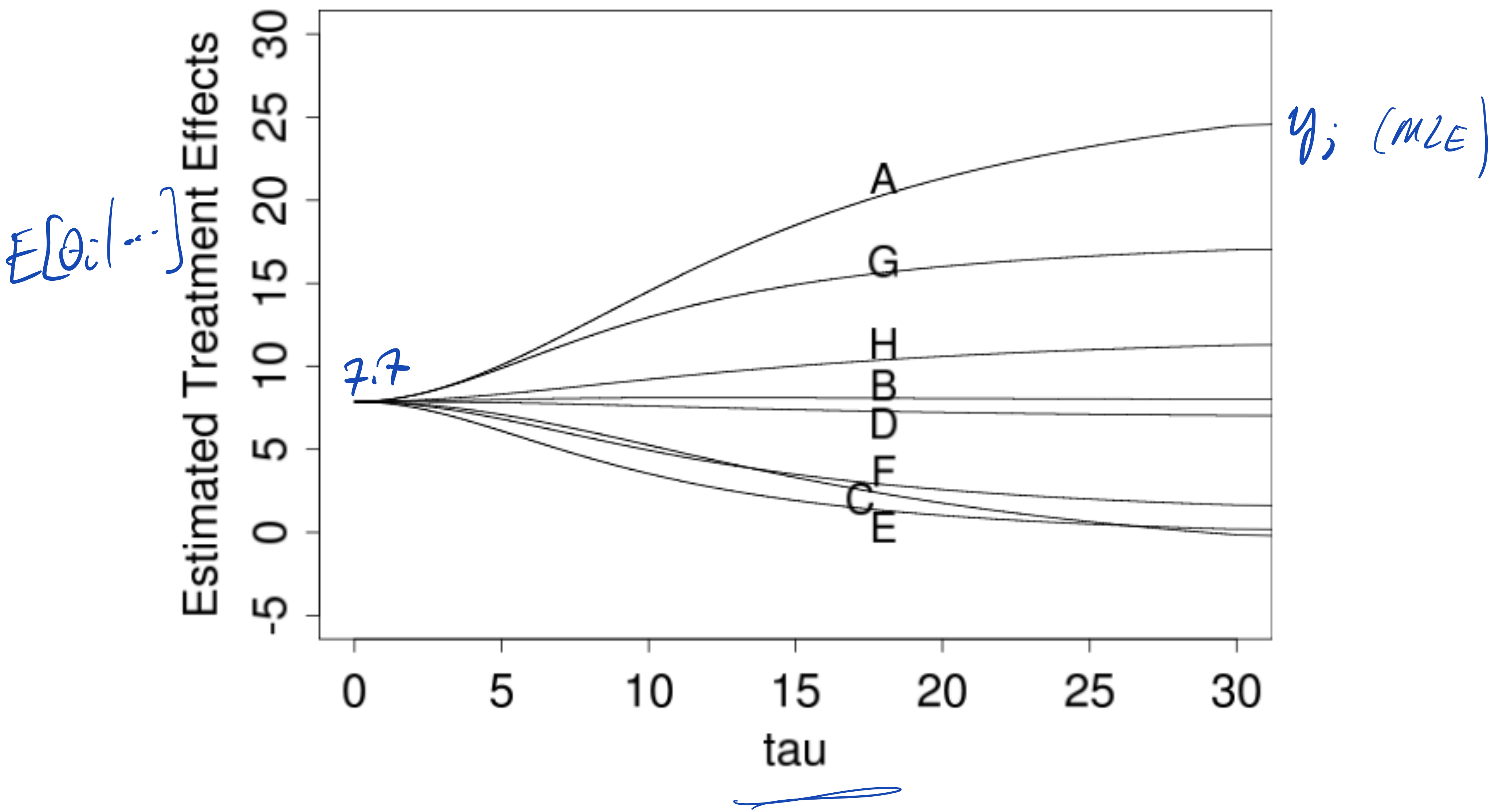
$$N(\sum \alpha y, \ \cdot)$$

Eight Schools example

$\theta_i \sim N(\mu, 1)$

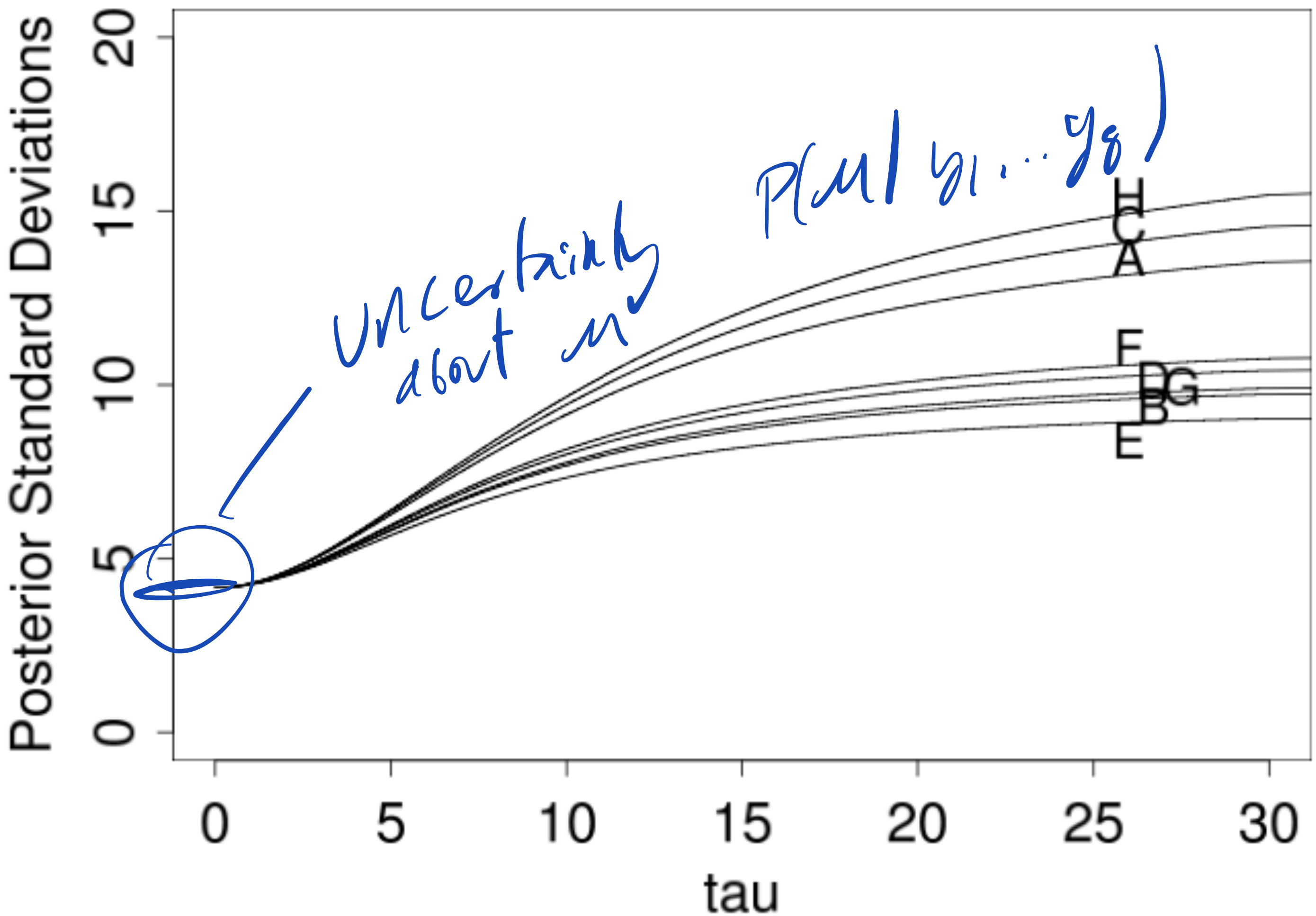


The impact of τ



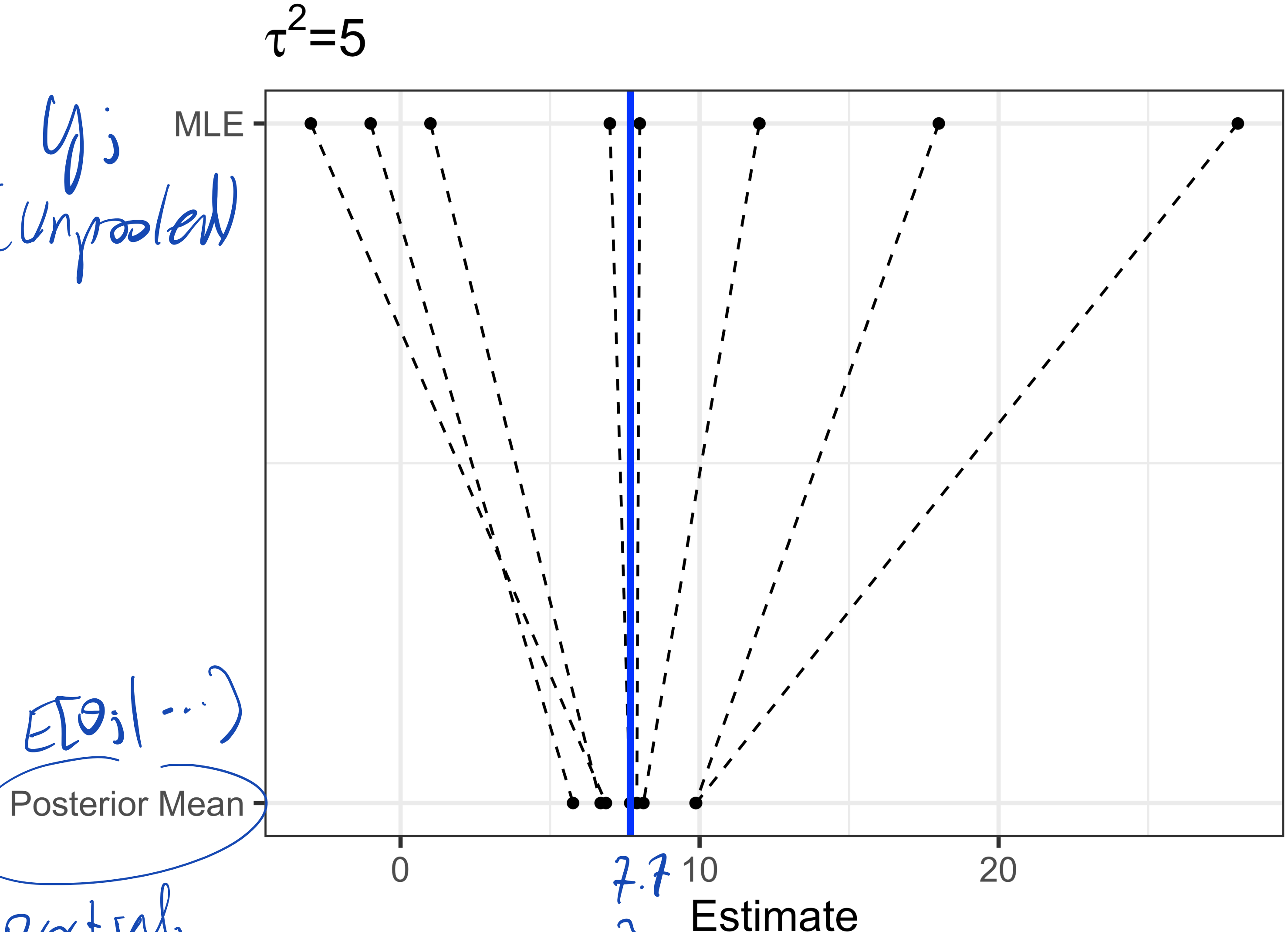
The impact of τ

Length of C.I.



MLE vs Posterior Mean

y_j (unpooled)



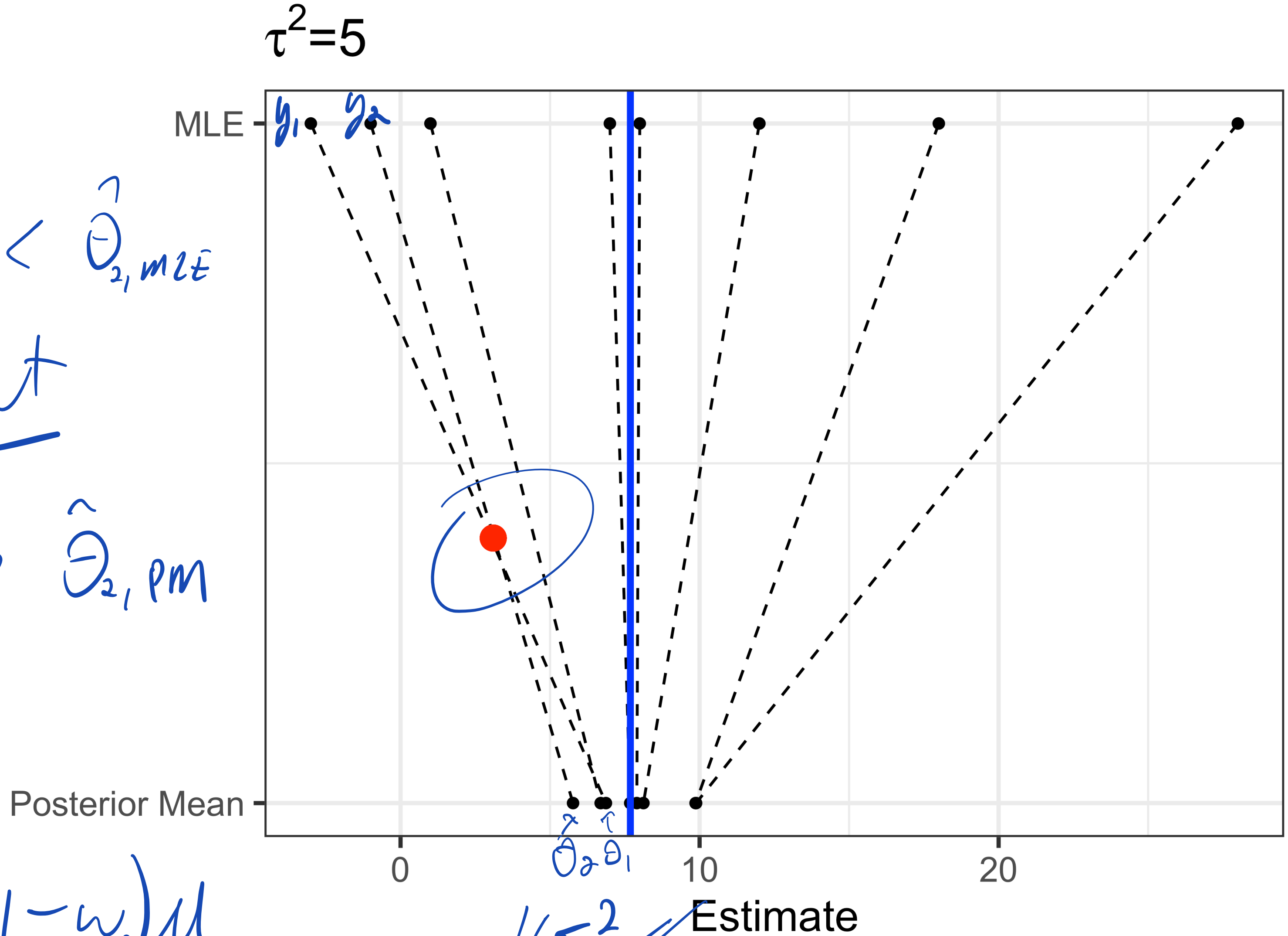
$E[\theta_j | \cdots)$

partial pooling.

7.7

$\hat{\mu}_{MLE}$

MLE vs Posterior Mean



$\hat{\theta}_{1,MLE} < \hat{\theta}_{2,MLE}$

But

$\hat{\theta}_{1,PM} > \hat{\theta}_{2,PM}$

$w_i y + (1-w_i)\mu$

$w = \dfrac{1/\sigma_i^2}{1/\sigma_i^2 + 1/\tau^2}$

Inference for τ^2

- Can infer τ^2 (don't need to set τ^2 as a hyperparameter)
- How?

Choose Prior for (μ, τ^2)

$\rightarrow P(\mu, \tau^2, \theta_1, \ldots \theta_8 \mid \ldots)$

Need a prior for τ^2.

$$\sigma_j^2 = \sigma^2$$

$$y_j \sim N(\theta_j, \sigma^2) \quad \text{(cond. on } \theta \text{ / cond } \mu)$$

$$\theta_j \sim N(\mu, \tau^2)$$

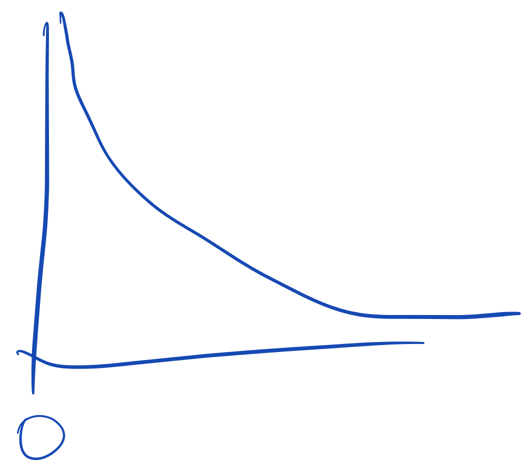
$$y_j \overset{iid.}{\sim} N(\mu, \sigma^2 + \tau^2) \quad \text{(uncond. } \theta \text{ / cond } \mu)$$

known.

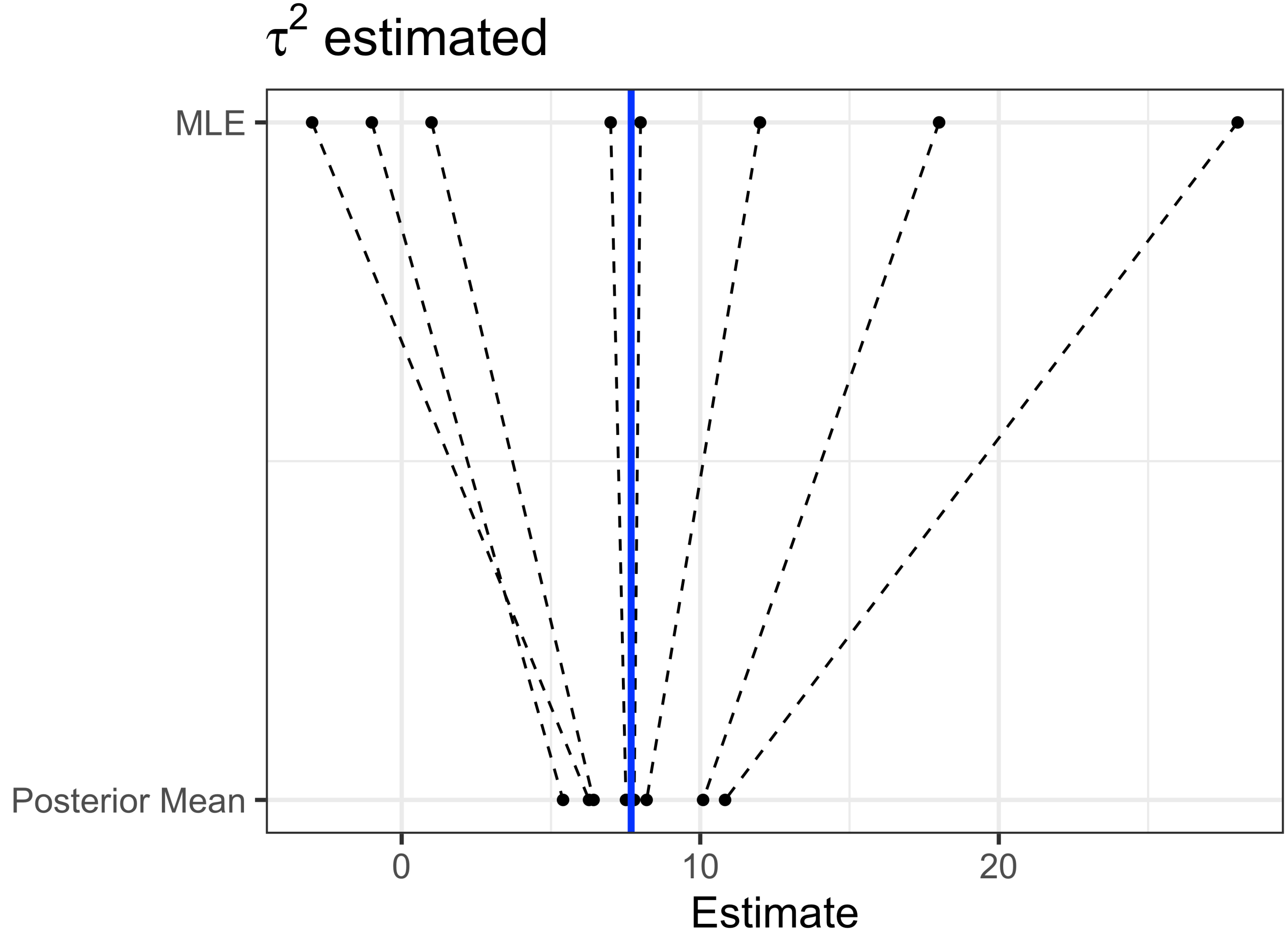
Weak and noninformative Priors on τ

Read Book

- Consider limits of proper priors
  - Uniform[0, A] as $A \to \infty$ (ok for $J > 2$)
  - Inverse-Gamma(ϵ, ϵ) as $\epsilon -> 0$ (improper posterior!)
- Uniform on log(τ) (improper posterior)
  - $p(y \mid \tau) \to const$ as $\tau \to 0$
- Half-Cauchy prior distribution on τ^2
  - Recommended by Gelman et al



MLE vs Posterior Mean



Posterior mean of $\tau = 5.6422886$