5: Hierarchical models

Taylor

University of Virginia

Introduction

On the one hand, we can assume that our data are iid, conditional on one parameter. On the other hand, we can assume that each data point gets its own parameter. The former might be too inflexible, while the latter might be too flexible, leading too overfitting.

Hierarchical models are agood "in between" option that allows each data point to get its own parameter; however, these parameters are "tied together" in a certain sense.

Rat tumor example

1. $j = 1, 2, \ldots, 71$ groups/experiments
2. θ_j is the probability of any rat getting a tumor in experiment j
3. θ_j are all different because of rat and/or experimental differences
4. y_j is the count of rats with tumors in experiment j (out of n_j total rats)
5. $y_j \mid \theta_j, n_j \sim \text{Binomial}(n_j, \theta_j)$ exchangeable
6. $\theta_j \overset{\text{iid}}{\sim} \text{Beta}(\alpha, \beta)$

Rat tumor example

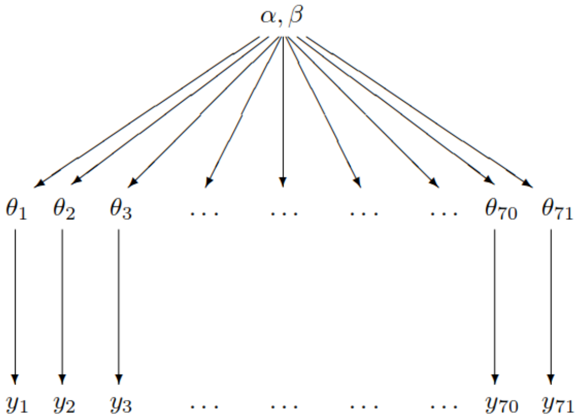


Figure 5.1: *Structure of the hierarchical model for the rat tumor example.*

Rat tumor example

Consider groups $1, \ldots, 70$ as historical data. We are interested specifically in θ_{71}.

Naive approach: only choose Beta(α, β) prior for θ_{71}. Choose α, β based on historical data $y_1, \ldots, y_{71}$, but in an ad hoc way, by setting the prior mean to be the empirical mean, and the prior variance equal to the sample variance. I.e. by solving

$$\begin{bmatrix} \hat{p} = 70^{-1} \sum_{i=1}^{70} y_j/n_j \\ 70^{-1} \sum_{i=1}^{70} (y_j/n_j - \hat{p})^2 \end{bmatrix} = \begin{bmatrix} \alpha/(\alpha+\beta) \\ \alpha\beta/\{(\alpha+\beta)^2(\alpha+\beta+1)\} \end{bmatrix}$$

You end up with $(\alpha, \beta) = (1.4, 8.6)$. Then, because $(y_{71}, n_{71}) = (4, 14)$,

$$\theta_{71} \mid y_{71} \sim \text{Beta}(5.4, 18.6)$$

Rat tumor example

Problems with this approach: .

1. can't really make inferences on $\theta_1, \ldots, \theta_{70}$ unless you "use the data twice"
2. how do we know we used the right point estimates for prior construction?
3. put a prior on α, β using prior knowledge, don't estimate with data and then assume these estimates are the "known" values (ignores uncertainty)

Rat tumor example

A better way:

1. choose (hyper)prior $p(\alpha, \beta)$
2. choose prior $p(\theta_{1:71} \mid \alpha, \beta)$
3. choose likelihood $p(y \mid \theta_{1:71}, \alpha, \beta) = p(y \mid \theta_{1:71}) = \prod_{j=1}^{71} p(y_j \mid \theta_j)$

Then

$$
\begin{aligned}
p(\theta_{1:71}, \alpha, \beta \mid y) &\propto p(y \mid \theta_{1:71}, \alpha, \beta) p(\theta_{1:71} \mid \alpha, \beta) p(\alpha, \beta) && \text{(Bayes')} \\
&= p(y \mid \theta_{1:71}) p(\theta_{1:71} \mid \alpha, \beta) p(\alpha, \beta) && \text{(condtl. indep.)}
\end{aligned}
$$

Posterior Predictive Distributions: two choices

If you want the predictive distribution of new rat counts ($\tilde{y}_{54}$) in an old/existing experiment (say $j = 54$), then you can use

$$p(\tilde{y}_{54} \mid y) = \int p(\tilde{y} \mid \theta_{54}) p(\theta_{54} \mid y) \mathrm{d}\theta_{54}$$

If you want the probability distribution of future rat counts ($\tilde{y}_{72}$) in a future experiment (say $j = 72$) coming from the same "superpopulation", you can use

$$p(\tilde{y}_{72} \mid y) = \iiint p(\tilde{y}_{72} \mid \theta_{72}) p(\theta_{72} \mid \alpha, \beta) p(\alpha, \beta \mid y) \mathrm{d}\theta_{72} \mathrm{d}\alpha \mathrm{d}\beta$$

Both strategies are based on the same decomposition, but the second way simulates twice.

Exchangeability in the prior

Why do we use exchangeable priors?

Q: If someone told you $\theta_1 = .2, \theta_2 = .3$, would you react differently than if they told you $\theta_2 = .2, \theta_1 = .3$?

Exchangeability in the prior

Why do we use exchangeable priors?

Q: If someone told you $\theta_1 = .2, \theta_2 = .3$, would you react differently than if they told you $\theta_2 = .2, \theta_1 = .3$?

A1: "No, I don't know anything about these labs, so it's all the same to me."
This means $p(\theta_{1:71})$ should be chosen to be exchangeable.

A2: "Yes, the second one is rarer a priori. I think θ_1 should be higher because the first lab sources their rats from NYC subways, and the second sources theirs from DC subways."

This means $p(\theta_{1:71})$ should not be chosen to be exchangeable.

Exchangeability in the prior

Let's say we assume exchangeability. How can we pick a prior?

Option 1: iid (not a hierarchical model)

$$p(\theta_{1:71}) = \prod_{i=1}^{71} p(\theta_i)$$

Does your opinion about θ_1 change if we knew θ_2? If yes, this isn't appropriate.

Option 2: mixture of iids

$$p(\theta_{1:71}) = \iint p(\theta_{1:71} \mid \alpha, \beta) p(\alpha, \beta) \mathrm{d}\alpha \mathrm{d}\beta = \iint \prod_{i=1}^{71} p(\theta_i \mid \alpha, \beta) p(\alpha, \beta) \mathrm{d}\alpha \mathrm{d}\beta$$

E.g. θ_1 and θ_2 are positively correlated (see HW2 #9).

Finding $p(\alpha, \beta \mid y)$

We choose the prior $p(\theta_{1:71} \mid \alpha, \beta)p(\alpha, \beta)$. 5.3 is mostly interested in the marginal posterior $p(\alpha, \beta \mid y)$. They advocate the following approach:

1. determine the conditional posterior in *closed form* $p(\theta_{1:71} \mid y, \alpha, \beta)$. This is only possible if you pick a **conditionally conjugate** $p(\theta_{1:71} \mid \alpha, \beta)$.

Finding $p(\alpha, \beta \mid y)$

We choose the prior $p(\theta_{1:71} \mid \alpha, \beta)p(\alpha, \beta)$. 5.3 is mostly interested in the marginal posterior $p(\alpha, \beta \mid y)$. They advocate the following approach:

1. determine the conditional posterior in *closed form* $p(\theta_{1:71} \mid y, \alpha, \beta)$. This is only possible if you pick a **conditionally conjugate** $p(\theta_{1:71} \mid \alpha, \beta)$.

2. determine the *unnormalized* version of the marginal posterior using the following formula

$$\begin{aligned} p(\alpha, \beta \mid y) &= p(\alpha, \beta \mid y)\frac{p(\theta_{1:71} \mid y, \alpha, \beta)}{p(\theta_{1:71} \mid y, \alpha, \beta)} \\ &= \frac{p(\theta_{1:71}, \alpha, \beta \mid y)}{p(\theta_{1:71} \mid y, \alpha, \beta)} \\ &\propto \frac{p(y \mid \theta_{1:71})p(\theta_{1:71} \mid \alpha, \beta)p(\alpha, \beta)}{p(\theta_{1:71} \mid y, \alpha, \beta)} \end{aligned}$$

Finding $p(\alpha, \beta \mid y)$

1. determine the conditional posterior in *closed form* $p(\theta_{1:71} \mid y, \alpha, \beta)$. We assume $\theta_{1:71} \mid \alpha, \beta \sim \text{Beta}(\alpha, \beta)$. Reminder: when we write $\propto$ we can drop anything that isn't a $\theta_{1:71}$.

$$
\begin{aligned}
p(\theta_{1:71} \mid y, \alpha, \beta) &\propto p(y \mid \theta_{1:71}) p(\theta_{1:71} \mid \alpha, \beta) \\
&\propto \prod_{j=1}^{71} \theta_j^{\alpha-1}(1-\theta_j)^{\beta-1} \\
&\times \prod_{j=1}^{71} \theta_j^{y_j}(1-\theta_j)^{n_j-y_j} \\
&= \prod_{j=1}^{71} \theta_j^{y_j+\alpha-1}(1-\theta_j)^{n_j-y_j+\beta-1}
\end{aligned}
$$

So $p(\theta_{1:71} \mid y, \alpha, \beta) = \prod_{j=1}^{71} \text{Beta}(\alpha + y_j, \beta + n_j - y - j)$

Finding $p(\alpha, \beta \mid y)$

2. determine the *unnormalized* version of the marginal posterior using the following formula. We use $p(\alpha, \beta) \propto (\alpha + \beta)^{-5/2}$. When we write $\propto$, we can drop anything that isn't involving α, β.

$$
\begin{aligned}
p(\alpha, \beta \mid y) &\propto \frac{p(y \mid \theta_{1:71}) p(\theta_{1:71} \mid \alpha, \beta) p(\alpha, \beta)}{p(\theta_{1:71} \mid y, \alpha, \beta)} && \text{(earlier slides)} \\
&\propto \frac{p(\theta_{1:71} \mid \alpha, \beta) p(\alpha, \beta)}{\prod_{j=1}^{71} \frac{\Gamma(n_j+\alpha+\beta)}{\Gamma(y_j+\alpha)\Gamma(n_j-y_j+\beta)} \theta_j^{y_j+\alpha-1}(1-\theta_j)^{n_j-y_j+\beta-1}} \\
&= \frac{\prod_{j=1}^{71} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta_j^{\alpha-1}(1-\theta_j)^{\beta-1}(\alpha+\beta)^{-5/2}}{\prod_{j=1}^{71} \frac{\Gamma(n_j+\alpha+\beta)}{\Gamma(y_j+\alpha)\Gamma(n_j-y_j+\beta)} \theta_j^{y_j+\alpha-1}(1-\theta_j)^{n_j-y_j+\beta-1}} \\
&\propto \prod_{j=1}^{71} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \Bigg/ \frac{\Gamma(n_j+\alpha+\beta)}{\Gamma(y_j+\alpha)\Gamma(n_j-y_j+\beta)}
\end{aligned}
$$

Finding $p(\alpha, \beta \mid y)$

$$p(\alpha, \beta \mid y) \propto \prod_{j=1}^{71} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \Big/ \frac{\Gamma(n_j+\alpha+\beta)}{\Gamma(y_j+\alpha)\Gamma(n_j-y_j+\beta)}$$

so

$$\log p(\alpha, \beta \mid y) = c + \sum_{j=1}^{71} \left\{ \log \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} - \log \frac{\Gamma(n_j+\alpha+\beta)}{\Gamma(y_j+\alpha)\Gamma(n_j-y_j+\beta)} \right\}$$

Finding $p(\alpha, \beta \mid y)$

```
A <- seq(0.5, 6, length.out = 100)
B <- seq(3, 33, length.out = 100)
cA <- rep(A, each = length(B))
cB <- rep(B, length(A))
lpfun <- function(a, b, y, n) log(a+b)*(-5/2) +
  sum(lgamma(a+b)-lgamma(a)-lgamma(b)+lgamma(a+y)+lgamma(b+n-y
lp <- mapply(lpfun, cA, cB, MoreArgs = list(y, n))
```

http:
//avehtari.github.io/BDA_R_demos/demos_ch5/demo5_1.html

Finding $p(\alpha, \beta \mid y)$

So then we exponentiate. But watch out:

```
> head(lp)
[1] -747.6954 -747.6320 -747.8540 -748.3062 -748.9466 -749.742
> head(exp(lp))
[1] 0 0 0 0 0 0
```

Finding $p(\alpha, \beta \mid y)$

So then we exponentiate. But watch out:

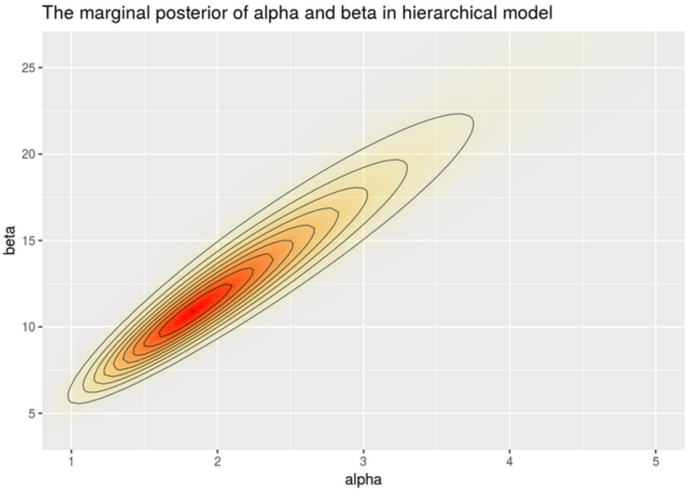
```
> head(lp)
[1] -747.6954 -747.6320 -747.8540 -748.3062 -748.9466 -749.742
> head(exp(lp))
[1] 0 0 0 0 0 0
```

This is **numerical underflow**. Solution:

$$p(\alpha, \beta \mid y) \propto \exp[\log p(\alpha, \beta \mid y) + m]$$

`m` is any "big" number. Careful not to set it too large, because then you will get **overflow**. The author uses a good data-dependent solution: set `m` to be equal to be negative of the maximum of these log-values, which is $\log p(\alpha, \beta \mid y)$.

Finding $p(\alpha, \beta \mid y)$



from http://avehtari.github.io/BDA_R_demos/demos_ch5/demo5_1.html

Finding $p(\alpha, \beta \mid y)$

Note that this plot is of the *unnormalized* marginal density. We do not know the normalizing constant!

They approximate the normalized density by making $p(\alpha, \beta \mid y)$ a discrete distribution defined on a grid (recall `A` and `B` from our code above). They evaluate the unnormalized density on this grid, but because there are a finite number of points, they can divide by the sum, yielding a pmf that sums to 1.

```
samp_indices <- sample(length(df_marg$p),
                       size = nsamp,
                       replace = T,
                       prob = df_marg$p/sum(df_marg$p))
```

Finding $p(\alpha, \beta \mid y)$

Note that this plot is of the *unnormalized* marginal density. We do not know the normalizing constant!

They approximate the normalized density by making $p(\alpha, \beta \mid y)$ a discrete distribution defined on a grid (recall `A` and `B` from our code above). They evaluate the unnormalized density on this grid, but because there are a finite number of points, they can divide by the sum, yielding a pmf that sums to 1.

```
samp_indices <- sample(length(df_marg$p),
                       size = nsamp,
                       replace = T,
                       prob = df_marg$p/sum(df_marg$p))
```

We will spend a lot of time in this class talking about different ways to sample from a posterior without knowing its normalizing constant!

Finding $p(\alpha, \beta \mid y)$

NB1: Anytime we have an improper prior, we must check that the posterior is proper! Our Bayesian inference will not make sense if it isn't. See question HW2 question 9.

Finding $p(\alpha, \beta \mid y)$

NB1: Anytime we have an improper prior, we must check that the posterior is proper! Our Bayesian inference will not make sense if it isn't. See question HW2 question 9.

NB2: You can do these calculations in the original parameter space (α, β), or you can do them in the transformed space $(\log(\alpha/\beta), \log(\alpha + \beta))$. If you choose the second one, you must use Jacobians for any distribution of α, β, but you do not need Jacobians for distributions that condition on α, β.

Finding $p(\alpha, \beta \mid y)$

NB1: Anytime we have an improper prior, we must check that the posterior is proper! Our Bayesian inference will not make sense if it isn't. See question HW2 question 9.

NB2: You can do these calculations in the original parameter space (α, β), or you can do them in the transformed space $(\log(\alpha/\beta), \log(\alpha + \beta))$. If you choose the second one, you must use Jacobians for any distribution of α, β, but you do not need Jacobians for distributions that condition on α, β.

NB3: You can see from pictures in the text, the posterior is less "pinched" in the transformed space. When we study MCMC algorithms, we will learn why pinchedness is undesirable.

Example 2: Normal Hierarchical Models

Normal hierarchical models aka one-way normal random-effects models, assume the following:

1. $j = 1, 2, \ldots, 71$ groups/experiments
2. $i = 1, \ldots, n_j$ replicates/observations for each experiment/group
3. $y_{i,j} \mid \theta_j \overset{\text{iid}}{\sim} \text{Normal}(\theta_j, \sigma^2)$
4. $\theta_j \mid \mu, \tau \overset{\text{iid}}{\sim} \text{Normal}(\mu, \tau)$
5. $\mu, \tau \sim p(\mu, \tau)$

Example 2: Normal Hierarchical Models

For each group j, because σ^2 is known, it's cleaner to replace

$$
\begin{aligned}
p(y_{1,j}, \ldots, y_{n_j,j} \mid \theta_j) &= \prod_{i=1}^{n_j} p(y_{i,j} \mid \theta_j) \\
&= (2\pi\sigma^2)^{-n_j/2} \exp\left[-\frac{1}{2\sigma^2}\sum_i (y_{i,j} - \theta_j)^2\right] \\
&= (2\pi\sigma^2)^{-n_j/2} \\
&\qquad \exp\left[-\frac{1}{2\sigma^2}\left\{\sum_i (y_{i,j} - \bar{y}_{\cdot,j})^2 + n_j(\bar{y}_{\cdot,j} - \theta_j)^2\right\}\right]
\end{aligned}
$$

with

$$
p(\bar{y}_{\cdot,j} \mid \theta_j) = (2\pi\sigma_j^2)^{-1/2} \exp\left[-\frac{1}{2\sigma_j^2}(\bar{y}_{\cdot,j} - \theta_j)^2\right]
$$

where $\bar{y}_{\cdot,j} = n_j^{-1}\sum_{i=1}^{n_j} y_{i,j}$ and $\sigma_j^2 = \sigma^2/n_j$

The joint posterior

We're going for everything this time. We want draws from the joint posterior distribution.

The primary decomposition:

$$p(\theta_{1:J}, \mu, \tau \mid y) = p(\tau \mid y)p(\mu \mid \tau, y)p(\theta_{1:J} \mid \mu, \tau, y)$$

The joint posterior

We're going for everything this time. We want draws from the joint posterior distribution.

The primary decomposition:

$$p(\theta_{1:J}, \mu, \tau \mid y) = p(\tau \mid y)p(\mu \mid \tau, y)p(\theta_{1:J} \mid \mu, \tau, y)$$

Do the following over and over again:

1. draw $\tau \sim p(\tau \mid y)$
2. draw $\mu \sim p(\mu \mid \tau, y)$
3. draw $\theta_{1:J} \sim p(\theta_{1:J} \mid \mu, \tau, y)$

We can prove the last two are normal so simulation is easy! Unfortunately, however, we're going to approximately draw from the first.

A conditional posterior

Result 1

$$p(\theta_{1:J} \mid \mu, \tau, y) = \prod_j \text{Normal}(\hat{\theta}_j, V_j)$$

where $V_j = \left(\frac{1}{\sigma_j^2} + \frac{1}{\tau^2} \right)^{-1}$ and $\hat{\theta}_j = V_j \left(\frac{1}{\sigma_j^2} \bar{y}_{\cdot,j} + \frac{1}{\tau^2} \mu \right)$

A conditional posterior

Proof:
Again, $p(\theta_{1:J} \mid \mu, \tau)$ is **conditionally conjugate**.

$$p(\theta_{1:J} \mid \mu, \tau, y) \propto \prod_{j=1}^{J} p(\bar{y}_{\cdot,j} \mid \theta_j) p(\theta_j \mid \mu, \tau)$$

Looking at each product:

$$\begin{aligned} p(\bar{y}_{\cdot,j} \mid \theta_j) p(\theta_j \mid \mu, \tau) &\propto \exp\left[-\frac{1}{2\sigma_j^2}\left(\bar{y}_{\cdot,j} - \theta_j\right)^2\right] \exp\left[-\frac{1}{2\tau^2}\left(\theta_j - \mu\right)^2\right] \\ &\propto \exp\left[-\frac{1}{2V_j}(\theta_j - \hat{\theta}_j)^2\right] \end{aligned}$$

where $V_j = \left(\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}\right)^{-1}$ and $\hat{\theta}_j = V_j\left(\frac{1}{\sigma_j^2}\bar{y}_{\cdot,j} + \frac{1}{\tau^2}\mu\right)$

Another conditional posterior

Result 2

If $p(\mu \mid \tau) \propto 1$, then

$$p(\mu \mid \tau, y) = \text{Normal}(\hat{\mu}, V_\mu)$$

where $V_\mu^{-1} = \sum_j \left(\frac{1}{\sigma_j^2 + \tau^2}\right)$ and $\hat{\mu} = V_\mu \sum_j \bar{y}_{\cdot,j} \left(\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}\right)$

Proof: homework question! Start by showing that $p(y \mid \mu, \tau) = \text{Normal}(\mu, \sigma_j^2 + \tau^2)$ (which we use again, later)

Another conditional posterior

Result 3

If $p(\tau) \propto 1$ and $p(\mu \mid \tau) \propto 1$, then

$$p(\tau \mid y) \propto V_\mu^{1/2} \prod_j (\sigma_j^2 + \tau^2)^{-1/2} \exp\left[-\frac{(\bar{y}_{\cdot,j} - \hat{\mu})^2}{2(\sigma_j^2 + \tau^2)}\right]$$

Another conditional posterior

Recall:

1. $p(y \mid \mu, \tau) = \prod_j \text{Normal}(\mu, \sigma_j^2 + \tau^2)$
2. $p(\mu \mid \tau, y) = \text{Normal}(\hat{\mu}, V_\mu)$

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

Another conditional posterior

$$p(\tau \mid y) \propto \frac{p(y \mid \mu, \tau)}{p(\mu \mid \tau, y)}$$

$$\propto \frac{\prod_j (\sigma_j^2 + \tau^2)^{-1/2} \exp\left[-\frac{1}{2(\sigma_j^2+\tau^2)}(\bar{y}_{\cdot,j} - \mu)^2\right]}{V_\mu^{-1/2} \exp\left[-\frac{1}{2V_\mu}(\mu - \hat{\mu})^2\right]}$$

$$\propto \frac{\prod_j (\sigma_j^2 + \tau^2)^{-1/2} \exp\left[-\frac{1}{2(\sigma_j^2+\tau^2)}(\mu - \hat{\mu} + \hat{\mu} - \bar{y}_{\cdot,j})^2\right]}{V_\mu^{-1/2} \exp\left[-\frac{1}{2V_\mu}(\mu - \hat{\mu})^2\right]}$$

$$\propto V_\mu^{1/2} \prod_j (\sigma_j^2 + \tau^2)^{-1/2} \exp\left[-\frac{(\bar{y}_{\cdot,j} - \hat{\mu})^2}{2(\sigma_j^2 + \tau^2)}\right]$$

last step is tricky!

Another conditional posterior

$$-\sum_j \frac{1}{(\sigma_j^2+\tau^2)}(\mu-\hat{\mu}+\hat{\mu}-\bar{y}_{\cdot,j})^2+\frac{1}{V_\mu}(\mu-\hat{\mu})^2$$

$$=\overbrace{-\sum_j \frac{(\hat{\mu}-\bar{y}_{\cdot,j})^2}{(\sigma_j^2+\tau^2)}}^{keep}-\sum_j \frac{(\mu-\hat{\mu})^2}{(\sigma_j^2+\tau^2)}+\sum_j \frac{2(\hat{\mu}-\bar{y}_{\cdot,j})(\mu-\hat{\mu})}{(\sigma_j^2+\tau^2)}+\frac{1}{V_\mu}(\mu-\hat{\mu})^2$$

$$=\overbrace{-\sum_j \frac{(\hat{\mu}-\bar{y}_{\cdot,j})^2}{(\sigma_j^2+\tau^2)}}^{keep}+2(\mu-\hat{\mu})\sum_j \frac{(\hat{\mu}-\bar{y}_{\cdot,j})}{(\sigma_j^2+\tau^2)}$$

$$=\overbrace{-\sum_j \frac{(\hat{\mu}-\bar{y}_{\cdot,j})^2}{(\sigma_j^2+\tau^2)}}^{keep}+2(\mu-\hat{\mu})(\hat{\mu}V_\mu^{-1}-\hat{\mu}V_\mu^{-1})=-\sum_j \frac{(\hat{\mu}-\bar{y}_{\cdot,j})^2}{(\sigma_j^2+\tau^2)}$$

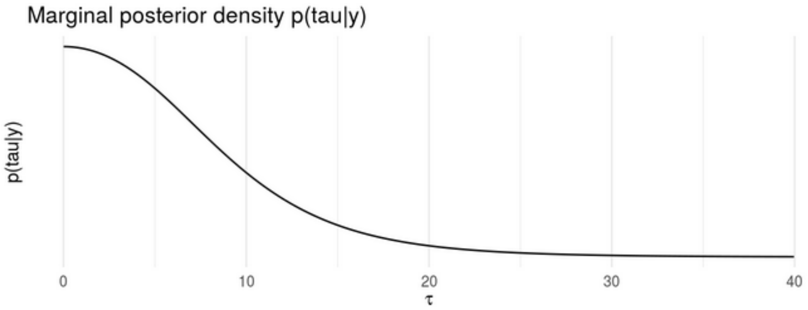
Example : how effective are SAT-V prep courses?

1. $j = 1, 2, \ldots, 8$ groups/experiments/schools
2. $\bar{y}_{\cdot,j}$ school j's effectiveness/treatment response measured in number of points
3. σ_j^2 is the (assumed) known variance of each of these
4. θ_j is the true effectiveness for each school

The book writes $\bar{y}_{\cdot,j}$ as y_j, but we use the first notation.

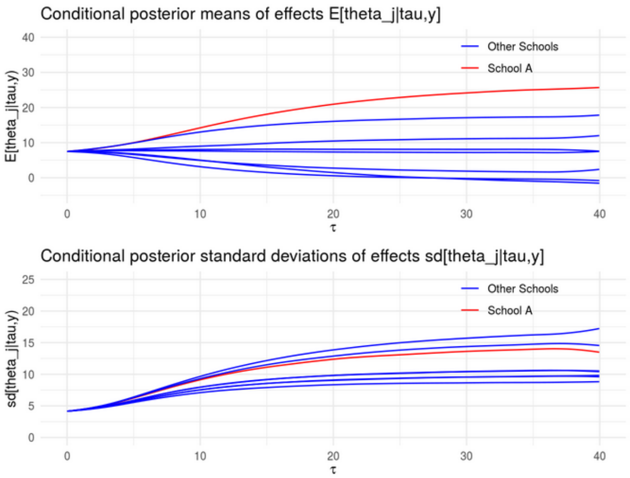
Example : how effective are SAT-V prep courses?

Goal 1: evaluate $p(\tau \mid y)$ on a grid



http://avehtari.github.io/BDA_R_demos/demos_ch5/demo5_2.html

Goal 2: summarize each $p(\theta_j \mid \tau, y) = \int p(\theta_j \mid \mu, \tau, y) p(\mu \mid \tau, y) d\mu$



avehtari.github.io/BDA_R_demos/demos_ch5/demo5_2.html