Foundations of Statistical Inference

J. Berestycki & D. Sejdinovic

Department of Statistics
University of Oxford

MT 2019

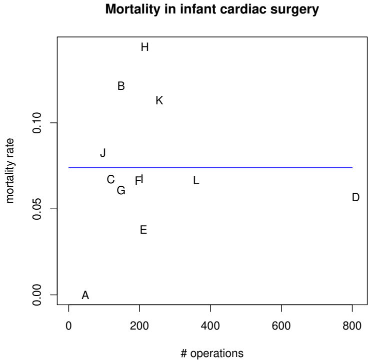
Chapter 10: Hierarchical Models

Basic idea

- The need to capture structure beyond what a single prior distribution on model parameters.
- Hierarchical modes: view parameters of a prior distribution as random variables that can be estimated from data.
- Motivation comes from joint inference on multiple parameters $\{\theta_1, \ldots, \theta_I\}$ *which are related or connected by the structure of the problem, but not identical*.

Example

Data from neonatal cardiac surgery in 12 hospitals. The number of operations in hospital i is n_i and the number of mortalities is y_i.



	A	B	C	D	E	F	G	H	I	J	K	L	$\sum$
y_i	0	18	8	46	8	13	9	31	14	8	29	24	208
n_i	47	148	119	810	211	196	148	215	207	97	256	360	2814

Three approaches

- **Identical parameters:** All the θ's are identical, in which case all the data can be pooled and the individual units ignored.
- **Independent parameters:** All the θ's are entirely unrelated, in which case the results from each unit can be analysed independently - individual estimates of θ_i are likely to be highly variable.
- **Exchangeable parameters:** The θ's are assumed to be 'similar' in the sense that the 'labels' convey no information.

ML estimates

1. the number of deaths $Y_i \sim \text{Bin}(n_i, \theta)$, ML estimate for *all* hospitals: $\hat{\theta} = \frac{\sum_i y_i}{\sum_i n_i} = 0.0739$.
   - Could the θ_i's all be equal? Variability in y_i suggests that this is not the case. For example, a test of $H_0 : \theta_H = 0.0739$ would reject at level $\alpha \ll 0.05/12$.
2. the number of deaths $Y_i \sim \text{Bin}(n_i, \theta_i)$, ML estimates: $\hat{\theta}_A = 0$, $\hat{\theta}_H = 0.1442$.
   - Should we really ignore data from all other hospitals when estimating θ_H? What if y_H was missing?

Different but related parameters

Allow for a different failure probability θ_i for each hospital i, but let θ_i come from the same distribution.

$$(y_i \mid \theta_i) \sim \text{Binomial}(n_i, \theta_i) \quad \text{where} \quad \theta_i \sim \text{Beta}(\alpha, \beta)$$

- But how would we specify the values for α and β?
- Say $\alpha = 4$, $\beta = 46$ (roughly "empirical Bayes" values), one obtains: $\hat{\theta}_A = 0.0412$, $\hat{\theta}_H = 0.1321$
- Bayesian estimates are 'pushed' towards the prior mean $\alpha/(\alpha + \beta) = 0.08$, to an extent depending on the 'denominator' n_i.

Empirical Bayes

- Calculate crude failure rates y_i/n_i
- Calculate the sample mean and variance of the 12 values y_i/n_i
- Solve for $\widehat{\alpha}$ and $\widehat{\beta}$ to obtain a beta distribution with this mean and variance
- Using Beta($\widehat{\alpha}$, $\widehat{\beta}$) as a prior, apply Bayes theorem to obtain posteriors for true failure rates θ_i, $p(\theta_i|\widehat{\alpha}, \widehat{\beta}, y_1, y_2, \ldots, y_I)$
  - uses the same data twice - overestimating precision
  - just one choice of (α, β) - ignoring uncertainty

Hierarchical Bayes

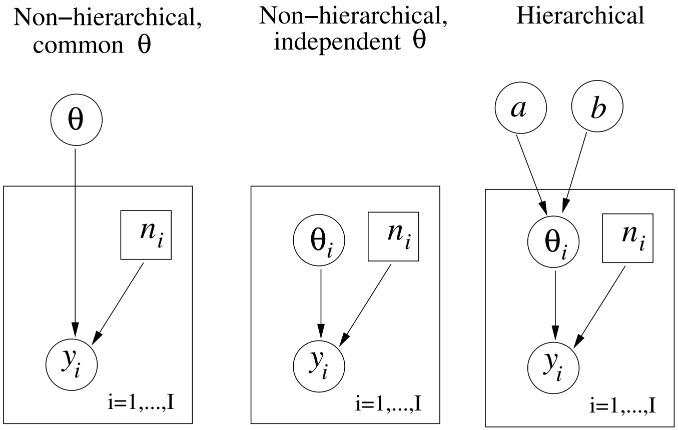
- Assume a *joint probability model* for the entire set of parameters (θ, α, β)
- Assign known prior distribution $\pi(\alpha, \beta)$ to α,β.
- Apply Bayes theorem to calculate the joint posterior distribution of all the unknown quantities simultaneously.

Level 1: $y_i \sim Binomial(n_i, \theta_i)$, independently for each i

Level 2: $\theta_i \sim Beta(\alpha, \beta)$, independently for each i

Level 3: hyperprior $\pi(\alpha, \beta)$

Hierarchical Bayes



Hierarchical Bayes

- Hierarchical modelling requires specification of conditional distributions which is natural in Bayesian approaches.
- Typical setting involves J experiments, observations $y_1, \dots, y_J$ with likelihoods $p(y_j|\theta_j)$.
- A full probabilistic model for the θ_j's is require. If the data is symmetric (i.e. there is no order on the experiments), then natural to assume that the distribution of the vector $(\theta_1, \dots, \theta_J)$ is symetric, i.e. exchangeable (invariant under relabelling by a permutation).

Exchangeability

- Symmetry among model parameters in the prior - invariant to permutations of the indices.
- When no information available to distinguish model parameters.
- True if drawn independently from a common distribution governed by a (hyper)parameter:

$$p(\theta_1, \theta_2, \ldots, \theta_I) = \int p(\phi) \prod_{i=1}^{I} p(\theta_i|\phi) d\phi.$$

- *converse* - De Finetti's theorem

Exchangeability

Definition

A sequence of random variables $(Y_1, \cdots, Y_n)$ is called exchangeable iff for all permutation σ of $\{1, \cdots, n\}$

$$(Y_{\sigma(1)}, \cdots, Y_{\sigma(n)}) \stackrel{\mathcal{D}}{=} (Y_1, \cdots, Y_n). \tag{1}$$

An infinite sequence of random variables $\{Y_i\}_{i\in\mathbb{N}}$ is called exchangeable iff (1) is true for all $n \in \mathbb{N}$.

Theorem (De Finetti)

An infinite sequence $\{Y_i\}_i$ is exchangeable iff there exists a random probability distribution P such that :

- *Conditionally on P, $\{Y_i\}_i | P \stackrel{i.i.d.}{\sim} P$*
- $P \sim \Pi$

Exchangeability

In the setting of the J experiments:

- If
$$Y_j|\theta_j \overset{ind}{\sim} f_j(Y_j|\theta_j), \quad \theta_j \overset{i.i.d}{\sim} G$$
The $(Y_j)_j$ are not exchangeable
- If
$$Y_j|\theta_j \overset{ind}{\sim} f(Y_j|\theta_j), \quad \theta_j \overset{i.i.d}{\sim} G \tag{2}$$
The $(Y_j)_j$ are exchangeable

(2) is the hierarchical representation of the model

$$f(y_1, \ldots, y_J|G) = \int_\Theta g(\theta) \prod_{j=1}^{J} f(y_j|\theta) d\theta.$$

Exchangeability

In the setting of the J experiments:

- If
$$Y_j|\theta_j \overset{ind}{\sim} f_j(Y_j|\theta_j), \quad \theta_j \overset{i.i.d}{\sim} G$$
The $(Y_j)_j$ are not exchangeable
- If
$$Y_j|\theta_j \overset{ind}{\sim} f(Y_j|\theta_j), \quad \theta_j \overset{i.i.d}{\sim} G \tag{2}$$
The $(Y_j)_j$ are exchangeable

(2) is the hierarchical representation of the model

$$f(y_1, \ldots, y_J|G) = \int_\Theta g(\theta) \prod_{j=1}^{J} f(y_j|\theta) d\theta.$$

Components of a hierarchical model

The prior has parameters which again have a probability distribution.

- Data y have a density $f(y|\theta)$. (In example : $y \sim B(n, \theta)$)
- The prior dist. of θ is $p(\theta|\psi)$. (In example: $\psi = (\alpha, \beta)$ and $\theta \sim Beta(\psi)$)
- ψ has a prior distribution $g(\psi)$, for $\psi \in \Psi$. New

Hierarchical model

- Joint prior: $p(\theta, \psi) = p(\theta|\psi)g(\psi)$
- Joint posterior: $p(\theta, \psi|y) \propto f(y|\theta)p(\theta|\psi)g(\psi)$,
- θ prior: $p(\theta) = \int p(\theta|\psi)g(\psi)d\psi$
- θ posterior $p(\theta|y) = \int_\Psi p(\theta, \psi|y)d\psi \propto f(y|\theta)p(\theta)$

Components of a hierarchical model

The prior has parameters which again have a probability distribution.

- Data y have a density $f(y|\theta)$. (In example : $y \sim B(n, \theta)$)
- The prior dist. of θ is $p(\theta|\psi)$. (In example: $\psi = (\alpha, \beta)$ and $\theta \sim Beta(\psi)$)
- ψ has a prior distribution $g(\psi)$, for $\psi \in \Psi$. New

Hierarchical model

- Joint prior: $p(\theta, \psi) = p(\theta|\psi)g(\psi)$
- Joint posterior: $p(\theta, \psi|y) \propto f(y|\theta)p(\theta|\psi)g(\psi)$,
- θ prior: $p(\theta) = \int p(\theta|\psi)g(\psi)d\psi$
- θ posterior $p(\theta|y) = \int_\Psi p(\theta, \psi|y)d\psi \propto f(y|\theta)p(\theta)$

Components of a hierarchical model

The prior has parameters which again have a probability distribution.

- Data y have a density $f(y|\theta)$. (In example : $y \sim B(n, \theta)$)
- The prior dist. of θ is $p(\theta|\psi)$. (In example: $\psi = (\alpha, \beta)$ and $\theta \sim Beta(\psi)$)
- ψ has a prior distribution $g(\psi)$, for $\psi \in \Psi$. New

Hierarchical model

- Joint prior: $p(\theta, \psi) = p(\theta|\psi)g(\psi)$
- Joint posterior: $p(\theta, \psi|y) \propto f(y|\theta)p(\theta|\psi)g(\psi)$,
- θ prior: $p(\theta) = \int p(\theta|\psi)g(\psi)d\psi$
- θ posterior $p(\theta|y) = \int_\Psi p(\theta, \psi|y)d\psi \propto f(y|\theta)p(\theta)$

Components of a hierarchical model

The prior has parameters which again have a probability distribution.

- Data y have a density $f(y|\theta)$. (In example : $y \sim B(n, \theta)$)
- The prior dist. of θ is $p(\theta|\psi)$. (In example: $\psi = (\alpha, \beta)$ and $\theta \sim Beta(\psi)$)
- ψ has a prior distribution $g(\psi)$, for $\psi \in \Psi$. New

Hierarchical model

- Joint prior: $p(\theta, \psi) = p(\theta|\psi)g(\psi)$
- Joint posterior: $p(\theta, \psi|y) \propto f(y|\theta)p(\theta|\psi)g(\psi)$,
- θ prior: $p(\theta) = \int p(\theta|\psi)g(\psi)d\psi$
- θ posterior $p(\theta|y) = \int_\Psi p(\theta, \psi|y)d\psi \propto f(y|\theta)p(\theta)$

Components of a hierarchical model

The prior has parameters which again have a probability distribution.

- Data y have a density $f(y|\theta)$. (In example : $y \sim B(n, \theta)$)
- The prior dist. of θ is $p(\theta|\psi)$. (In example: $\psi = (\alpha, \beta)$ and $\theta \sim Beta(\psi)$)
- ψ has a prior distribution $g(\psi)$, for $\psi \in \Psi$. New

Hierarchical model

- Joint prior: $p(\theta, \psi) = p(\theta|\psi)g(\psi)$
- Joint posterior: $p(\theta, \psi|y) \propto f(y|\theta)p(\theta|\psi)g(\psi)$,
- θ prior: $p(\theta) = \int p(\theta|\psi)g(\psi)d\psi$
- θ posterior $p(\theta|y) = \int_\Psi p(\theta, \psi|y)d\psi \propto f(y|\theta)p(\theta)$

Components of a hierarchical model

The prior has parameters which again have a probability distribution.

- Data y have a density $f(y|\theta)$. (In example : $y \sim B(n, \theta)$)
- The prior dist. of θ is $p(\theta|\psi)$. (In example: $\psi = (\alpha, \beta)$ and $\theta \sim Beta(\psi)$)
- ψ has a prior distribution $g(\psi)$, for $\psi \in \Psi$. New

Hierarchical model

- Joint prior: $p(\theta, \psi) = p(\theta|\psi)g(\psi)$
- Joint posterior: $p(\theta, \psi|y) \propto f(y|\theta)p(\theta|\psi)g(\psi)$,
- θ prior: $p(\theta) = \int p(\theta|\psi)g(\psi)d\psi$
- θ posterior $p(\theta|y) = \int_\Psi p(\theta, \psi|y)d\psi \propto f(y|\theta)p(\theta)$

Components of a hierarchical model

The prior has parameters which again have a probability distribution.

- Data y have a density $f(y|\theta)$. (In example : $y \sim B(n, \theta)$)
- The prior dist. of θ is $p(\theta|\psi)$. (In example: $\psi = (\alpha, \beta)$ and $\theta \sim Beta(\psi)$)
- ψ has a prior distribution $g(\psi)$, for $\psi \in \Psi$. New

Hierarchical model

- Joint prior: $p(\theta, \psi) = p(\theta|\psi)g(\psi)$
- Joint posterior: $p(\theta, \psi|y) \propto f(y|\theta)p(\theta|\psi)g(\psi)$,
- θ prior: $p(\theta) = \int p(\theta|\psi)g(\psi)d\psi$
- θ posterior $p(\theta|y) = \int_\Psi p(\theta, \psi|y)d\psi \propto f(y|\theta)p(\theta)$

Components of a hierarchical model

The prior has parameters which again have a probability distribution.

- Data y have a density $f(y|\theta)$. (In example : $y \sim B(n, \theta)$)
- The prior dist. of θ is $p(\theta|\psi)$. (In example: $\psi = (\alpha, \beta)$ and $\theta \sim Beta(\psi)$)
- ψ has a prior distribution $g(\psi)$, for $\psi \in \Psi$. New

Hierarchical model

- Joint prior: $p(\theta, \psi) = p(\theta|\psi)g(\psi)$
- Joint posterior: $p(\theta, \psi|y) \propto f(y|\theta)p(\theta|\psi)g(\psi)$,
- θ prior: $p(\theta) = \int p(\theta|\psi)g(\psi)d\psi$
- θ posterior $p(\theta|y) = \int_\Psi p(\theta, \psi|y)d\psi \propto f(y|\theta)p(\theta)$

Components of a hierarchical model

The prior has parameters which again have a probability distribution.

- Data y have a density $f(y|\theta)$. (In example : $y \sim B(n, \theta)$)
- The prior dist. of θ is $p(\theta|\psi)$. (In example: $\psi = (\alpha, \beta)$ and $\theta \sim Beta(\psi)$)
- ψ has a prior distribution $g(\psi)$, for $\psi \in \Psi$. New

Hierarchical model

- Joint prior: $p(\theta, \psi) = p(\theta|\psi)g(\psi)$
- Joint posterior: $p(\theta, \psi|y) \propto f(y|\theta)p(\theta|\psi)g(\psi)$,
- θ prior: $p(\theta) = \int p(\theta|\psi)g(\psi)d\psi$
- θ posterior $p(\theta|y) = \int_\Psi p(\theta, \psi|y)d\psi \propto f(y|\theta)p(\theta)$

Components of a hierarchical model

The prior has parameters which again have a probability distribution.

- Data y have a density $f(y|\theta)$. (In example : $y \sim B(n, \theta)$)
- The prior dist. of θ is $p(\theta|\psi)$. (In example: $\psi = (\alpha, \beta)$ and $\theta \sim Beta(\psi)$)
- ψ has a prior distribution $g(\psi)$, for $\psi \in \Psi$. New

Hierarchical model

- Joint prior: $p(\theta, \psi) = p(\theta|\psi)g(\psi)$
- Joint posterior: $p(\theta, \psi|y) \propto f(y|\theta)p(\theta|\psi)g(\psi)$,
- θ prior: $p(\theta) = \int p(\theta|\psi)g(\psi)d\psi$
- θ posterior $p(\theta|y) = \int_\Psi p(\theta, \psi|y)d\psi \propto f(y|\theta)p(\theta)$

Analysis of a hierarchical model

To analyze a hierarchical model :

1. Write the joint posterior $p(\theta, \psi|y)$, in unnormalized form as the product $p(y|\theta) \times p(\theta|\psi) \times g(\psi)$.
2. Determine $p(\theta|\psi, y)$ (the conditional posterior density of θ given ψ for fixed observation y.
3. Obtain $p(\psi|y)$ the posterior marginal distribution of hyper parameter ψ given the observation y.

For the last step, observe that

$$p(\psi|y) = \frac{p(\theta, \psi|y)}{p(\theta|\psi, y)}.$$

Analysis of a hierarchical model

To analyze a hierarchical model :

1. Write the joint posterior $p(\theta, \psi|y)$, in unnormalized form as the product $p(y|\theta) \times p(\theta|\psi) \times g(\psi)$.
2. Determine $p(\theta|\psi, y)$ (the conditional posterior density of θ given ψ for fixed observation y.
3. Obtain $p(\psi|y)$ the posterior marginal distribution of hyper parameter ψ given the observation y.

For the last step, observe that

$$p(\psi|y) = \frac{p(\theta, \psi|y)}{p(\theta|\psi, y)}.$$

Analysis of a hierarchical model

To analyze a hierarchical model :

1. Write the joint posterior $p(\theta, \psi|y)$, in unnormalized form as the product $p(y|\theta) \times p(\theta|\psi) \times g(\psi)$.
2. Determine $p(\theta|\psi, y)$ (the conditional posterior density of θ given ψ for fixed observation y.
3. Obtain $p(\psi|y)$ the posterior marginal distribution of hyper parameter ψ given the observation y.

For the last step, observe that

$$p(\psi|y) = \frac{p(\theta, \psi|y)}{p(\theta|\psi, y)}.$$

Analysis of a hierarchical model

To analyze a hierarchical model :

1. Write the joint posterior $p(\theta, \psi|y)$, in unnormalized form as the product $p(y|\theta) \times p(\theta|\psi) \times g(\psi)$.
2. Determine $p(\theta|\psi, y)$ (the conditional posterior density of θ given ψ for fixed observation y.
3. Obtain $p(\psi|y)$ the posterior marginal distribution of hyper parameter ψ given the observation y.

For the last step, observe that

$$p(\psi|y) = \frac{p(\theta, \psi|y)}{p(\theta|\psi, y)}.$$

Analysis of a hierarchical model

To analyze a hierarchical model :

1. Write the joint posterior $p(\theta, \psi|y)$, in unnormalized form as the product $p(y|\theta) \times p(\theta|\psi) \times g(\psi)$.
2. Determine $p(\theta|\psi, y)$ (the conditional posterior density of θ given ψ for fixed observation y.
3. Obtain $p(\psi|y)$ the posterior marginal distribution of hyper parameter ψ given the observation y.

For the last step, observe that

$$p(\psi|y) = \frac{p(\theta, \psi|y)}{p(\theta|\psi, y)}.$$

Analysis of a hierarchical model

To analyze a hierarchical model :

1. Write the joint posterior $p(\theta, \psi|y)$, in unnormalized form as the product $p(y|\theta) \times p(\theta|\psi) \times g(\psi)$. Immediate
2. Determine $p(\theta|\psi, y)$ (the conditional posterior density of θ given ψ for fixed observation y.
3. Obtain $p(\psi|y)$ the posterior marginal distribution of hyper parameter ψ given the observation y.

For the last step, observe that

$$p(\psi|y) = \frac{p(\theta, \psi|y)}{p(\theta|\psi, y)}.$$

Analysis of a hierarchical model

To analyze a hierarchical model :

1. Write the joint posterior $p(\theta, \psi|y)$, in unnormalized form as the product $p(y|\theta) \times p(\theta|\psi) \times g(\psi)$. Immediate
2. Determine $p(\theta|\psi, y)$ (the conditional posterior density of θ given ψ for fixed observation y. Easy for conjugate models since θ_j are iid cond. on ψ
3. Obtain $p(\psi|y)$ the posterior marginal distribution of hyper parameter ψ given the observation y.

For the last step, observe that

$$p(\psi|y) = \frac{p(\theta, \psi|y)}{p(\theta|\psi, y)}.$$

Analysis of a hierarchical model

To analyze a hierarchical model :

1. Write the joint posterior $p(\theta, \psi|y)$, in unnormalized form as the product $p(y|\theta) \times p(\theta|\psi) \times g(\psi)$. Immediate
2. Determine $p(\theta|\psi, y)$ (the conditional posterior density of θ given ψ for fixed observation y. Easy for conjugate models since θ_j are iid cond. on ψ
3. Obtain $p(\psi|y)$ the posterior marginal distribution of hyper parameter ψ given the observation y. Integrate joint posterior over θ

For the last step, observe that

$$p(\psi|y) = \frac{p(\theta, \psi|y)}{p(\theta|\psi, y)}.$$

Analysis of a hierarchical model

To analyze a hierarchical model :

1. Write the joint posterior $p(\theta, \psi|y)$, in unnormalized form as the product $p(y|\theta) \times p(\theta|\psi) \times g(\psi)$. Immediate
2. Determine $p(\theta|\psi, y)$ (the conditional posterior density of θ given ψ for fixed observation y. Easy for conjugate models since θ_j are iid cond. on ψ
3. Obtain $p(\psi|y)$ the posterior marginal distribution of hyper parameter ψ given the observation y. Integrate joint posterior over θ

For the last step, observe that

$$p(\psi|y) = \frac{p(\theta, \psi|y)}{p(\theta|\psi, y)}.$$

Careful about normalizing factor.

Example cont'd

Full probability model:

- the y_j are independent with $y_j \sim B(n_j, \theta_j)$.
- the θ_j are i.i.d. Beta(α, β)
- $\psi = (\alpha, \beta)$ follows an uninformative prior to be specified.

We now perform the three steps of the analysis.

Step1: Joint posterior

$$p(\theta, \alpha, \beta | y) \propto p(\alpha, \beta) p(\theta | \alpha, \beta) p(y | \theta, \alpha, \beta)$$
$$\propto p(\alpha, \beta) \prod_{j=1}^{J} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta_j^{\alpha-1} (1-\theta_j)^{\beta-1} \prod_{j=1}^{J} \theta_j^{y_j} (1-\theta_j)^{n_j - y_j}.$$

Step2: Posterior density of θ given (α, β)

$$p(\theta | \alpha, \beta, y) = \prod_{j=1}^{J} \frac{\Gamma(\alpha+\beta+n_j)}{\Gamma(\alpha+y_j)\Gamma(\beta+n_j-y_j)} \theta_j^{\alpha+y_j-1} (1-\theta_j)^{\beta+n_j-y_j-1}$$

Example cont'd

Full probability model:

- the y_j are independent with $y_j \sim B(n_j, \theta_j)$.
- the θ_j are i.i.d. Beta(α, β)
- $\psi = (\alpha, \beta)$ follows an uninformative prior to be specified.

We now perform the three steps of the analysis.

Step1: Joint posterior

$$\begin{aligned} p(\theta, \alpha, \beta | y) &\propto p(\alpha, \beta) p(\theta | \alpha, \beta) p(y | \theta, \alpha, \beta) \\ &\propto p(\alpha, \beta) \prod_{j=1}^{J} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta_j^{\alpha-1} (1-\theta_j)^{\beta-1} \prod_{j=1}^{J} \theta_j^{y_j} (1-\theta_j)^{n_j - y_j}. \end{aligned}$$

Step2: Posterior density of θ given (α, β)

$$p(\theta | \alpha, \beta, y) = \prod_{j=1}^{J} \frac{\Gamma(\alpha+\beta+n_j)}{\Gamma(\alpha+y_j)\Gamma(\beta+n_j-y_j)} \theta_j^{\alpha+y_j-1} (1-\theta_j)^{\beta+n_j-y_j-1}$$

Example cont'd

Full probability model:

- the y_j are independent with $y_j \sim B(n_j, \theta_j)$.
- the θ_j are i.i.d. Beta(α, β)
- $\psi = (\alpha, \beta)$ follows an uninformative prior to be specified.

We now perform the three steps of the analysis.

Step1: Joint posterior

$$
\begin{aligned}
p(\theta, \alpha, \beta | y) &\propto p(\alpha, \beta) p(\theta | \alpha, \beta) p(y | \theta, \alpha, \beta) \\
&\propto p(\alpha, \beta) \prod_{j=1}^{J} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta_j^{\alpha-1} (1-\theta_j)^{\beta-1} \prod_{j=1}^{J} \theta_j^{y_j} (1-\theta_j)^{n_j - y_j}.
\end{aligned}
$$

Step2: Posterior density of θ given (α, β)

$$
p(\theta | \alpha, \beta, y) = \prod_{j=1}^{J} \frac{\Gamma(\alpha+\beta+n_j)}{\Gamma(\alpha+y_j)\Gamma(\beta+n_j-y_j)} \theta_j^{\alpha+y_j-1} (1-\theta_j)^{\beta+n_j-y_j-1}
$$

Example cont'd

Step3: Posterior of α, β using $p(\phi|y) = p(\theta, \phi|y)/p(\theta|\phi, y)$

$$p(\alpha, \beta|y) \propto p(\alpha, \beta) \prod_{j=1}^{J} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+y_j)\Gamma(\beta+n_j-y_j)}{\Gamma(\alpha+\beta+n_j)}.$$

Possible noninformative prior for α, β

- Uniform in α, β: $p(\alpha, \beta) \propto 1$. Does it yield a proper posterior? No
- Recall that mean is $\alpha/(\alpha+\beta)$ and that $\alpha+\beta$ is 'sample size'. Take logit and log to put them on a $(-\infty, \infty)$ scale and then assign a uniform prior: $p(\log(\alpha/\beta), \log(\alpha+\beta)) \propto 1$ No
- A reasonable choice of diffuse hyperprior density is uniform on $(\alpha/(\alpha+\beta), (\alpha+\beta)^{-1/2})$ which translates to $p(\alpha, \beta) \propto (\alpha+\beta)^{-5/2}$, and yields a proper posterior.

Example cont'd

Step3: Posterior of α, β using $p(\phi|y) = p(\theta, \phi|y)/p(\theta|\phi, y)$

$$p(\alpha, \beta|y) \propto p(\alpha, \beta) \prod_{j=1}^{J} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+y_j)\Gamma(\beta+n_j-y_j)}{\Gamma(\alpha+\beta+n_j)}.$$

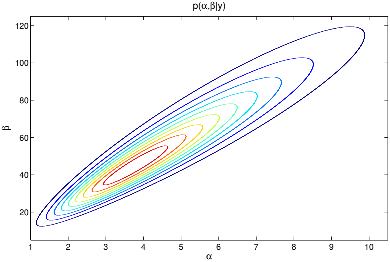
Possible noninformative prior for α, β

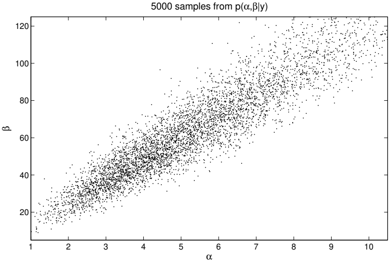
- Uniform in α, β: $p(\alpha, \beta) \propto 1$. Does it yield a proper posterior? No
- Recall that mean is $\alpha/(\alpha+\beta)$ and that $\alpha+\beta$ is 'sample size'. Take logit and $\log$ to put them on a $(-\infty, \infty)$ scale and then assign a uniform prior: $p(\log(\alpha/\beta), \log(\alpha+\beta)) \propto 1$ No
- A reasonable choice of diffuse hyperprior density is uniform on $(\alpha/(\alpha+\beta), (\alpha+\beta)^{-1/2})$ which translates to $p(\alpha, \beta) \propto (\alpha+\beta)^{-5/2}$, and yields a proper posterior.

Simulating from the posterior

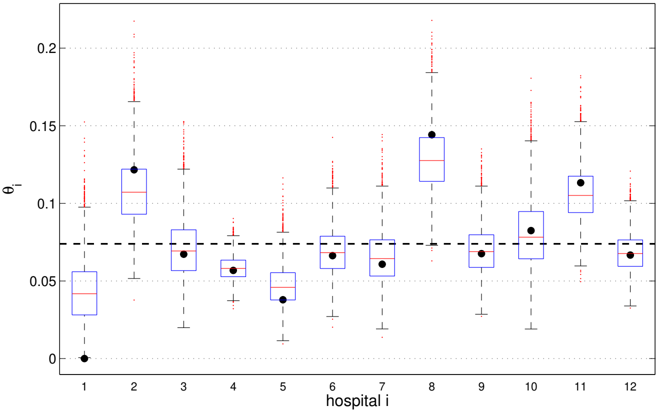
- Draw (α, β) from $p(\alpha, \beta|y)$
- Draw θ from $p(\theta|\alpha, \beta, y)$ given the drawn value of (α, β). Since $p(\theta|\alpha, \beta, y) = \prod_i p(\theta_i|\alpha, \beta, y)$, components θ_i can be drawn independently.
- Predictive values $\tilde{y}$ can be drawn from $p(\tilde{y}|\theta)$, given the drawn θ.

Posterior $p(\alpha, \beta|y)$





Posterior $p(\theta|y)$



Advantages

The posterior distribution for each θ_i

- *'borrows strength'* from the likelihood contributions for *all* hospitals, via their joint influence on the estimate of the unknown population (prior) parameters α and β
- reflects our full uncertainty about the true values of α and β

Example: Normal data – ANOVA type model

For $i = 1, 2, ..., J$ we make n_i observations $X_{i,1}, X_{i,2}, ..., X_{i,n_i}$ on population i, with $X_{ij} \sim N(\theta_i, \sigma^2)$. The θ_i are the unknown means for observations on the i'th population but σ^2 is known.

Question: what sort of estimates for θ given the (y_{ij})?

- Simple natural idea: $\hat{\theta}_j = \bar{y}_{\cdot j} = \frac{1}{n_j} \sum_{i=1}^{n_j} y_{ij}$
- If the J experiments are very close might prefer $\hat{\theta}_j = \hat{\theta} = \bar{y}_{\cdot\cdot} = \frac{1}{N} \sum_{i,j=1}^{n_j, J} y_{ij}$

To decide which to use, usually ANOVA F-test.

Example: Normal data – ANOVA type model

For $i = 1, 2, ..., J$ we make n_i observations $X_{i,1}, X_{i,2}, ..., X_{i,n_i}$ on population i, with $X_{ij} \sim N(\theta_i, \sigma^2)$. The θ_i are the unknown means for observations on the i'th population but σ^2 is known.

Question: what sort of estimates for θ given the (y_{ij})?

- Simple natural idea: $\hat{\theta}_j = \bar{y}_{\cdot j} = \frac{1}{n_j} \sum_{i=1}^{n_j} y_{ij}$
- If the J experiments are very close might prefer $\hat{\theta}_j = \hat{\theta} = \bar{y}_{\cdot\cdot} = \frac{1}{N} \sum_{i,j=1}^{n_j,J} y_{ij}$

To decide which to use, usually ANOVA F-test.

Example: Normal data – ANOVA type model

For $i = 1, 2, ..., J$ we make n_i observations $X_{i,1}, X_{i,2}, ..., X_{i,n_i}$ on population i, with $X_{ij} \sim N(\theta_i, \sigma^2)$. The θ_i are the unknown means for observations on the i'th population but σ^2 is known.
Question: what sort of estimates for θ given the (y_{ij})?

- Simple natural idea: $\hat{\theta}_j = \bar{y}_{.j} = \frac{1}{n_j} \sum_{i=1}^{n_j} y_{ij}$
- If the J experiments are very close might prefer $\hat{\theta}_j = \hat{\theta} = \bar{y}_{..} = \frac{1}{N} \sum_{i,j=1}^{n_j, J} y_{ij}$

To decide which to use, usually ANOVA F-test.

Example: Normal data

But we could also interpolate

$$\hat{\theta}_j = \lambda_j \bar{y}_{\cdot j} + (1 - \lambda_j)\bar{y}_{\cdot\cdot}.$$

1. The unpooled estimate $\hat{\theta}_j = \bar{y}_{\cdot j}, \lambda_j = 1$ corresponds to θ_j having independent uniform priors
2. The pooled estimate $\lambda_j = 0$ corresponds to the θ_j restricted to be equal with uniform prior.
3. The weighted estimates $\lambda_j \in (0, 1)$ corresponds to the case where the θ_j are iid normal.

Example: Normal data

But we could also interpolate

$$\hat{\theta}_j = \lambda_j \bar{y}_{\cdot j} + (1 - \lambda_j)\bar{y}_{\cdot\cdot}.$$

1. The unpooled estimate $\hat{\theta}_j = \bar{y}_{\cdot j}, \lambda_j = 1$ corresponds to θ_j having independent uniform priors
2. The pooled estimate $\lambda_j = 0$ corresponds to the θ_j restricted to be equal with uniform prior.
3. The weighted estimates $\lambda_j \in (0, 1)$ corresponds to the case where the θ_j are iid normal.

Example: Hierarchical model for normal data

For $i = 1, 2, ..., k$ we make n_i observations $X_{i,1}, X_{i,2}, ..., X_{i,n_i}$ on population i, with $X_{ij} \sim N(\theta_i, \sigma^2)$. The θ_i are the unknown means for observations on the i'th population but σ^2 is known. Suppose the prior model for the θ_i is iid normal, $\theta_i \sim N(\phi, \tau^2)$.

$$
\begin{array}{lcc}
X_{1,1}, \ldots, X_{1,n_1} \sim N(\theta_1, \sigma^2) & \theta_1 & \\
X_{2,1}, \ldots, X_{2,n_2} \sim N(\theta_2, \sigma^2) & \theta_2 & \\
\quad\quad\quad . & . & N(\phi, \tau^2) \\
\quad\quad\quad . & . & \\
X_{k,1}, \ldots, X_{k,n_k} \sim N(\theta_k, \sigma^2) & \theta_k &
\end{array}
$$

Example: Hierarchical model for normal data

For $i = 1, 2, ..., k$ we make n_i observations $X_{i,1}, X_{i,2}, ..., X_{i,n_i}$ on population i, with $X_{ij} \sim N(\theta_i, \sigma^2)$. The θ_i are the unknown means for observations on the i'th population but σ^2 is known. Suppose the prior model for the θ_i is iid normal, $\theta_i \sim N(\phi, \tau^2)$.

$$
\begin{array}{lcl}
X_{1,1}, \ldots, X_{1,n_1} \sim N\left(\theta_1, \sigma^2\right) & \qquad \theta_1 & \\
X_{2,1}, \ldots, X_{2,n_2} \sim N\left(\theta_2, \sigma^2\right) & \qquad \theta_2 & \\
\qquad\qquad . & \qquad . & \qquad N\left(\phi, \tau^2\right) \\
\qquad\qquad . & \qquad . & \\
X_{k,1}, \ldots, X_{k,n_k} \sim N\left(\theta_k, \sigma^2\right) & \qquad \theta_k &
\end{array}
$$

Example: Hierarchical model for normal data

If $\psi = (\phi, \tau^2)$

$$\pi(\theta_1, \ldots, \theta_k|\psi) = \prod_{i=1}^{k} (2\pi\tau^2)^{-1/2} \exp\left\{-\frac{1}{2\tau^2}(\theta_i - \phi)^2\right\},$$

Now we need a prior for ϕ and τ^2. Suppose we take

$$g(\phi, \tau^2) = p(\phi|\tau)p(\tau) \propto p(\tau),$$

i.e. ϕ is uniform conditionally on τ. Keep $p(\tau)$ for later.

The joint posterior of the parameters is

$$\begin{aligned}\pi(\theta, \psi|x) &\propto f(x;\theta)\pi(\theta|\psi)g(\psi) \\ &\propto g(\psi) \prod_{j=1}^{J} N(\theta_j|\phi, \tau^2) \prod_{j=1}^{J} N(\bar{y}_{\cdot j}|\theta_j, \sigma_j^2)\end{aligned}$$

where $\sigma_j^2 = \sigma^2/n_j$

Example: Hierarchical model for normal data

If $\psi = (\phi, \tau^2)$

$$\pi(\theta_1, \ldots, \theta_k | \psi) = \prod_{i=1}^{k} (2\pi\tau^2)^{-1/2} \exp\left\{ -\frac{1}{2\tau^2} (\theta_i - \phi)^2 \right\},$$

Now we need a prior for ϕ and τ^2. Suppose we take

$$g(\phi, \tau^2) = p(\phi|\tau) p(\tau) \propto p(\tau),$$

i.e. ϕ is uniform conditionally on τ. Keep $p(\tau)$ for later.

The joint posterior of the parameters is

$$\begin{aligned} \pi(\theta, \psi | x) &\propto f(x; \theta) \pi(\theta|\psi) g(\psi) \\ &\propto g(\psi) \prod_{j=1}^{J} N(\theta_j | \phi, \tau^2) \prod_{j=1}^{J} N(\bar{y}_{\cdot j} | \theta_j, \sigma_j^2) \end{aligned}$$

where $\sigma_j^2 = \sigma^2 / n_j$

Example: Hierarchical model for normal data

If $\psi = (\phi, \tau^2)$

$$\pi(\theta_1, \dots, \theta_k | \psi) = \prod_{i=1}^{k} (2\pi\tau^2)^{-1/2} \exp\left\{ -\frac{1}{2\tau^2} (\theta_i - \phi)^2 \right\},$$

Now we need a prior for ϕ and τ^2. Suppose we take

$$g(\phi, \tau^2) = p(\phi|\tau) p(\tau) \propto p(\tau),$$

i.e. ϕ is uniform conditionally on τ. Keep $p(\tau)$ for later.

The joint posterior of the parameters is

$$\begin{aligned}\pi(\theta, \psi | x) &\propto f(x; \theta) \pi(\theta|\psi) g(\psi) \\ &\propto g(\psi) \prod_{j=1}^{J} N(\theta_j | \phi, \tau^2) \prod_{j=1}^{J} N(\bar{y}_{\cdot j} | \theta_j, \sigma_j^2)\end{aligned}$$

where $\sigma_j^2 = \sigma^2 / n_j$

Example: Hierarchical model for normal data

Step 2: Now we want to fix ψ and write the conditional posterior of θ. Because conditionally on ψ the θ_j are iid we can treat each θ_j in turn

$$\theta_j|\phi, \tau^2, y \sim N(\hat{\theta}_j, V_j)$$

with

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

Step 3 Now we go full Bayesian on the hyperparameters.

$$p(\phi, \tau|y) \propto g(\phi, \tau)p(y|\phi, \tau).$$

In general this expression is no help because $p(y|\phi, \tau)$ doesn't have a closed form. But here

$$p(\phi, \tau|y) \propto g(\phi, \tau)\prod_{j=1}^{J} N(\bar{y}_{\cdot j}|\phi, \tau^2 + \sigma^2).$$

Example: Hierarchical model for normal data

Step 2: Now we want to fix ψ and write the conditional posterior of θ. Because conditionally on ψ the θ_j are iid we can treat each θ_j in turn

$$\theta_j|\phi, \tau^2, y \sim N(\hat{\theta}_j, V_j)$$

with

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

Step 3 Now we go full Bayesian on the hyperparameters.

$$p(\phi, \tau|y) \propto g(\phi, \tau)p(y|\phi, \tau).$$

In general this expression is no help because $p(y|\phi, \tau)$ doesn't have a closed form. But here

$$p(\phi, \tau|y) \propto g(\phi, \tau)\prod_{j=1}^{J} N(\bar{y}_{\cdot j}|\phi, \tau^2 + \sigma^2).$$

Example: Hierarchical model for normal data

Step 2: Now we want to fix ψ and write the conditional posterior of θ. Because conditionally on ψ the θ_j are iid we can treat each θ_j in turn

$$\theta_j|\phi, \tau^2, y \sim N(\hat{\theta}_j, V_j)$$

with

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

Step 3 Now we go full Bayesian on the hyperparameters.

$$p(\phi, \tau|y) \propto g(\phi, \tau)p(y|\phi, \tau).$$

In general this expression is no help because $p(y|\phi, \tau)$ doesn't have a closed form. But here

$$p(\phi, \tau|y) \propto g(\phi, \tau) \prod_{j=1}^{J} N(\bar{y}_{\cdot j}|\phi, \tau^2 + \sigma^2).$$

Example: Hierarchical model for normal data

$$p(\phi, \tau | y) \propto g(\phi, \tau) \prod_{j=1}^{J} N(\bar{y}_{\cdot j} | \phi, \tau^2 + \sigma^2).$$

Start by fixing τ and compute $p(\phi|\tau, y)$. Using that $g(\phi, \tau^2) \propto p(\tau)$ we see that $\log p(\phi|\tau, y)$ is quadratic in ϕ and thus

$$\phi|\tau, y \sim N(\hat{\phi}, V_\phi) \quad \text{where} \quad \hat{\phi} = \frac{\sum_{j=1}^{J} \frac{\bar{y}_{\cdot j}}{\sigma_j^2 + \tau^2}}{\sum_{j=1}^{J} \frac{1}{\sigma_j^2 + \tau^2}} \quad V_\phi^{-1} = \left(\sum_{j=1}^{J} \frac{1}{\sigma_j^2 + \tau^2} \right)^{-1}$$

This is a proper posterior for ϕ given τ.
Using $p(\phi, \tau|y) = p(\phi|\tau, y)p(\tau|y)$ we get

$$p(\tau|y) \propto \frac{p(\tau) \prod_{j=1}^{J} N(\bar{y}_{\cdot j}|\phi, \tau^2 + \sigma_j^2)}{N(\phi|\hat{\phi}, V_\phi)}$$

Example: Hierarchical model for normal data

$$p(\phi, \tau | y) \propto g(\phi, \tau) \prod_{j=1}^{J} N(\bar{y}_{\cdot j} | \phi, \tau^2 + \sigma^2).$$

Start by fixing τ and compute $p(\phi|\tau, y)$. Using that $g(\phi, \tau^2) \propto p(\tau)$ we see that $\log p(\phi|\tau, y)$ is quadratic in ϕ and thus

$$\phi|\tau, y \sim N(\hat{\phi}, V_\phi) \quad \text{where} \quad \hat{\phi} = \frac{\sum_{j=1}^{J} \frac{\bar{y}_{\cdot j}}{\sigma_j^2 + \tau^2}}{\sum_{j=1}^{J} \frac{1}{\sigma_j^2 + \tau^2}} \quad V_\phi^{-1} = \left(\sum_{j=1}^{J} \frac{1}{\sigma_j^2 + \tau^2} \right)^{-1}$$

This is a proper posterior for ϕ given τ.

Using $p(\phi, \tau | y) = p(\phi|\tau, y) p(\tau|y)$ we get

$$p(\tau|y) \propto \frac{p(\tau) \prod_{j=1}^{J} N(\bar{y}_{\cdot j} | \phi, \tau^2 + \sigma_j^2)}{N(\phi | \hat{\phi}, V_\phi)}$$

Example: Hierarchical model for normal data

$$p(\phi, \tau | y) \propto g(\phi, \tau) \prod_{j=1}^{J} N(\bar{y}_{\cdot j} | \phi, \tau^2 + \sigma^2).$$

Start by fixing τ and compute $p(\phi|\tau, y)$. Using that $g(\phi, \tau^2) \propto p(\tau)$ we see that $\log p(\phi|\tau, y)$ is quadratic in ϕ and thus

$$\phi | \tau, y \sim N(\hat{\phi}, V_\phi) \quad \text{where} \quad \hat{\phi} = \frac{\sum_{j=1}^{J} \frac{\bar{y}_{\cdot j}}{\sigma_j^2 + \tau^2}}{\sum_{j=1}^{J} \frac{1}{\sigma_j^2 + \tau^2}} \quad V_\phi^{-1} = \left(\sum_{j=1}^{J} \frac{1}{\sigma_j^2 + \tau^2} \right)^{-1}$$

This is a proper posterior for ϕ given τ.
Using $p(\phi, \tau | y) = p(\phi | \tau, y) p(\tau | y)$ we get

$$p(\tau | y) \propto \frac{p(\tau) \prod_{j=1}^{J} N(\bar{y}_{\cdot j} | \phi, \tau^2 + \sigma_j^2)}{N(\phi | \hat{\phi}, V_\phi)}$$

Example: Hierarchical model for normal data

Using $p(\phi, \tau|y) = p(\phi|\tau, y)p(\tau|y)$ we get

$$p(\tau|y) \propto \frac{p(\tau)\prod_{j=1}^{J} N(\bar{y}_{\cdot j}|\phi, \tau^2 + \sigma_j^2)}{N(\phi|\hat{\phi}, V_\phi)}$$

Trick: Must hold for any value of μ so all μ terms must simplify away. In particular, must hold for $\mu = \hat{\mu}$.

$$p(\tau|y) \propto p(\tau)V_\phi^{1/2}\prod_{j=1}^{J}(\tau^2 + \sigma_j^2)^{-1/2}\exp\left\{-\frac{(\bar{y}_{\cdot j} - \hat{\phi})^2}{2(\sigma_j^2 + \tau^2)}\right\}$$

Both $\hat{\phi}$ and V_ϕ are functions of τ.

Example: Hierarchical model for normal data

Using $p(\phi, \tau|y) = p(\phi|\tau, y)p(\tau|y)$ we get

$$p(\tau|y) \propto \frac{p(\tau)\prod_{j=1}^{J} N(\bar{y}_{\cdot j}|\phi, \tau^2 + \sigma_j^2)}{N(\phi|\hat{\phi}, V_\phi)}$$

Trick: Must hold for any value of μ so all μ terms must simplify away. In particular, must hold for $\mu = \hat{\mu}$.

$$p(\tau|y) \propto p(\tau)V_\phi^{1/2}\prod_{j=1}^{J}(\tau^2 + \sigma_j^2)^{-1/2}\exp\left\{-\frac{(\bar{y}_{\cdot j} - \hat{\phi})^2}{2(\sigma_j^2 + \tau^2)}\right\}$$

Both $\hat{\phi}$ and V_ϕ are functions of τ.

Example: Hierarchical model for normal data

Using $p(\phi, \tau|y) = p(\phi|\tau, y)p(\tau|y)$ we get

$$p(\tau|y) \propto \frac{p(\tau)\prod_{j=1}^{J} N(\bar{y}_{\cdot j}|\phi, \tau^2 + \sigma_j^2)}{N(\phi|\hat{\phi}, V_\phi)}$$

Trick: Must hold for any value of μ so all μ terms must simplify away. In particular, must hold for $\mu = \hat{\mu}$.

$$p(\tau|y) \propto p(\tau)V_\phi^{1/2}\prod_{j=1}^{J}(\tau^2 + \sigma_j^2)^{-1/2}\exp\left\{-\frac{(\bar{y}_{\cdot j} - \hat{\phi})^2}{2(\sigma_j^2 + \tau^2)}\right\}$$

Both $\hat{\phi}$ and V_ϕ are functions of τ.

Example: Hierarchical model for normal data

We now want the posterior of θ given the observations y.

Either

$$p(\theta|y) = \int p(\theta|y,(\phi,\tau))p(\phi,\tau|y)d\phi d\tau = \int \text{step 2} \times \text{step 3}$$

or

$$p(\theta|y) = \int p(\theta,(\phi,\tau)|y)d\phi d\tau$$

Using the second approach with $p(\tau) \propto \tau^{-a}$, $a \geq 0$

$$\begin{aligned} \pi(\theta,\phi,\tau^2|x) &\propto \left[\prod_{j=1}^{J} \exp\left\{-\frac{1}{2\sigma_j^2}(\bar{y}_{\cdot j}-\theta_j)^2\right\}\right] \\ &\times \tau^{-a}\left[\prod_{j=1}^{J} \tau^{-1}\exp\left\{-\frac{1}{2\tau^2}(\theta_j-\phi)^2\right\}\right] \end{aligned}$$

Example: Hierarchical model for normal data

We now want the posterior of θ given the observations y.
Either

$$p(\theta|y) = \int p(\theta|y,(\phi,\tau))p(\phi,\tau|y)d\phi d\tau = \int \text{step 2} \times \text{step 3}$$

or

$$p(\theta|y) = \int p(\theta,(\phi,\tau)|y)d\phi d\tau$$

Using the second approach with $p(\tau) \propto \tau^{-a}$, $a \geq 0$

$$\begin{aligned}
\pi(\theta,\phi,\tau^2|x) &\propto \left[\prod_{j=1}^{J} \exp\left\{-\frac{1}{2\sigma_j^2}(\bar{y}_{\cdot j} - \theta_j)^2\right\}\right] \\
&\times \tau^{-a}\left[\prod_{j=1}^{J} \tau^{-1}\exp\left\{-\frac{1}{2\tau^2}(\theta_j - \phi)^2\right\}\right]
\end{aligned}$$

Example: Hierarchical model for normal data

We now want the posterior of θ given the observations y.
Either

$$p(\theta|y) = \int p(\theta|y,(\phi,\tau))p(\phi,\tau|y)d\phi d\tau = \int \text{step 2} \times \text{step 3}$$

or

$$p(\theta|y) = \int p(\theta,(\phi,\tau)|y)d\phi d\tau$$

Using the second approach with $p(\tau) \propto \tau^{-a}$, $a \geq 0$

$$\begin{aligned}\pi(\theta,\phi,\tau^2|x) &\propto \left[\prod_{j=1}^{J}\exp\left\{-\frac{1}{2\sigma_j^2}(\bar{y}_{\cdot j}-\theta_j)^2\right\}\right] \\ &\times \tau^{-a}\left[\prod_{j=1}^{J}\tau^{-1}\exp\left\{-\frac{1}{2\tau^2}(\theta_j-\phi)^2\right\}\right]\end{aligned}$$

Example: Hierarchical model for normal data

We now want the posterior of θ given the observations y.
Either

$$p(\theta|y) = \int p(\theta|y,(\phi,\tau))p(\phi,\tau|y)d\phi d\tau = \int \text{step 2} \times \text{step 3}$$

or

$$p(\theta|y) = \int p(\theta,(\phi,\tau)|y)d\phi d\tau$$

Using the second approach with $p(\tau) \propto \tau^{-a}$, $a \geq 0$

$$\begin{aligned}\pi(\theta,\phi,\tau^2|x) &\propto \left[\prod_{j=1}^{J} \exp\left\{-\frac{1}{2\sigma_j^2}(\bar{y}_{\cdot j}-\theta_j)^2\right\}\right] \\ &\times \tau^{-a}\left[\prod_{j=1}^{J} \tau^{-1}\exp\left\{-\frac{1}{2\tau^2}(\theta_j-\phi)^2\right\}\right]\end{aligned}$$

Example: Hierarchical model for normal data

We now want the posterior of θ given the observations y.
Either

$$p(\theta|y) = \int p(\theta|y,(\phi,\tau))p(\phi,\tau|y)d\phi d\tau = \int \text{step 2} \times \text{step 3}$$

or

$$p(\theta|y) = \int p(\theta,(\phi,\tau)|y)d\phi d\tau$$

Using the second approach with $p(\tau) \propto \tau^{-a}$, $a \geq 0$

$$\begin{aligned}\pi(\theta,\phi,\tau^2|x) &\propto \left[\prod_{j=1}^{J} \exp\left\{-\frac{1}{2\sigma_j^2}(\bar{y}_{\cdot j}-\theta_j)^2\right\}\right] \\ &\times \tau^{-a}\left[\prod_{j=1}^{J} \tau^{-1}\exp\left\{-\frac{1}{2\tau^2}(\theta_j-\phi)^2\right\}\right]\end{aligned}$$

Example: Hierarchical model for normal data

We now want the posterior of θ given the observations y.
Either

$$p(\theta|y) = \int p(\theta|y,(\phi,\tau))p(\phi,\tau|y)d\phi d\tau = \int \text{step 2} \times \text{step 3}$$

or

$$p(\theta|y) = \int p(\theta,(\phi,\tau)|y)d\phi d\tau$$

Using the second approach with $p(\tau) \propto \tau^{-a}$, $a \geq 0$

$$\begin{aligned}
\pi(\theta,\phi,\tau^2|x) &\propto \left[\prod_{j=1}^{J} \exp\left\{-\frac{1}{2\sigma_j^2}(\bar{y}_{\cdot j}-\theta_j)^2\right\}\right] \\
&\times \tau^{-a}\left[\prod_{j=1}^{J} \tau^{-1}\exp\left\{-\frac{1}{2\tau^2}(\theta_j-\phi)^2\right\}\right]
\end{aligned}$$

Example: Hierarchical model for normal data

$$\begin{aligned}\pi(\theta,\phi,\tau^2|x) &\propto \left[\prod_{j=1}^{J}\exp\left\{-\frac{1}{2\sigma_j^2}(\bar{y}_{\cdot j}-\theta_j)^2\right\}\right]\\ &\times \tau^{-a}\left[\prod_{j=1}^{J}\tau^{-1}\exp\left\{-\frac{1}{2\tau^2}(\theta_j-\phi)^2\right\}\right]\end{aligned}$$

Integrate out wrt ϕ and τ^2 to obtain $\pi(\theta|x)$.

Exercise Integrating the last factor wrt ϕ gives a term proportional to

$$\tau^{1-J-a}\exp\left\{-\frac{1}{2\tau^2}\sum(\theta_j-\bar{\theta})^2\right\}.$$

Exercise Then the integral wrt τ gives a term proportional to

$$\left[\sum(\theta_j-\bar{\theta})^2\right]^{1-(J+a)/2}.$$

Example: Hierarchical model for normal data

$$\pi(\theta, \phi, \tau^2|x) \quad \propto \quad \left[\prod_{j=1}^{J} \exp\left\{-\frac{1}{2\sigma_j^2}(\bar{y}_{\cdot j} - \theta_j)^2\right\}\right]$$
$$\times \quad \tau^{-a}\left[\prod_{j=1}^{J} \tau^{-1}\exp\left\{-\frac{1}{2\tau^2}(\theta_j - \phi)^2\right\}\right]$$

Integrate out wrt ϕ and τ^2 to obtain $\pi(\theta|x)$.
Exercise Integrating the last factor wrt ϕ gives a term proportional to

$$\tau^{1-J-a}\exp\left\{-\frac{1}{2\tau^2}\sum(\theta_j - \bar{\theta})^2\right\}.$$

Exercise Then the integral wrt τ gives a term proportional to

$$\left[\sum(\theta_j - \bar{\theta})^2\right]^{1-(J+a)/2}.$$

Example: Hierarchical model for normal data

$$\begin{aligned}\pi(\theta, \phi, \tau^2|x) &\propto \left[\prod_{j=1}^{J} \exp\left\{-\frac{1}{2\sigma_j^2}(\bar{y}_{\cdot j} - \theta_j)^2\right\}\right] \\ &\times \tau^{-a}\left[\prod_{j=1}^{J} \tau^{-1} \exp\left\{-\frac{1}{2\tau^2}(\theta_j - \phi)^2\right\}\right]\end{aligned}$$

Integrate out wrt ϕ and τ^2 to obtain $\pi(\theta|x)$.
Exercise Integrating the last factor wrt ϕ gives a term proportional to

$$\tau^{1-J-a} \exp\left\{-\frac{1}{2\tau^2}\sum(\theta_j - \bar{\theta})^2\right\}.$$

Exercise Then the integral wrt τ gives a term proportional to

$$\left[\sum(\theta_j - \bar{\theta})^2\right]^{1-(J+a)/2}.$$

Example: Hierarchical model for normal data

Thus the posterior distribution of θ is

$$\pi(\theta|x) \quad \propto \quad \left[\prod_{i=1}^{J} \exp\left\{-\frac{1}{2\sigma_j^2}(\bar{y}_{\cdot j} - \theta_j)^2\right\}\right] \cdot \left[\sum (\theta_j - \bar{\theta})^2\right]^{1-(J+a)/2}$$

Integrable iff $J + a - 2 > J - 1$ iff $a > -1$.

If the θ_j were unrelated then $\hat{\theta}_j = \bar{y}_{\cdot j}$. The model modifies the estimate by pulling it towards the mean of the estimated θ_is.

Another kind of interpolation model.

Example: Hierarchical model for normal data

Thus the posterior distribution of θ is

$$\pi(\theta|x) \quad \propto \quad \left[\prod_{i=1}^{J} \exp\left\{-\frac{1}{2\sigma_j^2}(\bar{y}_{\cdot j} - \theta_j)^2\right\}\right] \cdot \left[\sum (\theta_j - \bar{\theta})^2\right]^{1-(J+a)/2}$$

Integrable iff $J + a - 2 > J - 1$ iff $a > -1$.

If the θ_j were unrelated then $\hat{\theta}_j = \bar{y}_{\cdot j}$. The model modifies the estimate by pulling it towards the mean of the estimated θ_is.
Another kind of interpolation model.