Foundations of Statistical Inference

Julien Berestycki

Department of Statistics
University of Oxford

MT 2016

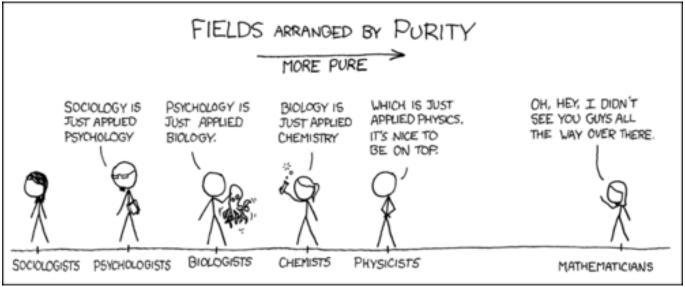
Lecture 8 : Hierarchical models. Bayesian Hypothesis Testing

Hierarchical models

Some models have a natural hierarchical structure.

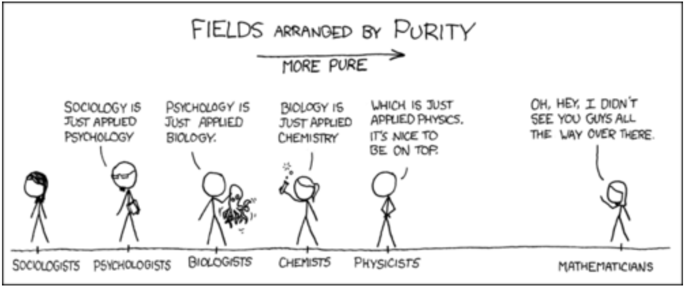
Hierarchical models

Some models have a natural hierarchical structure.



Hierarchical models

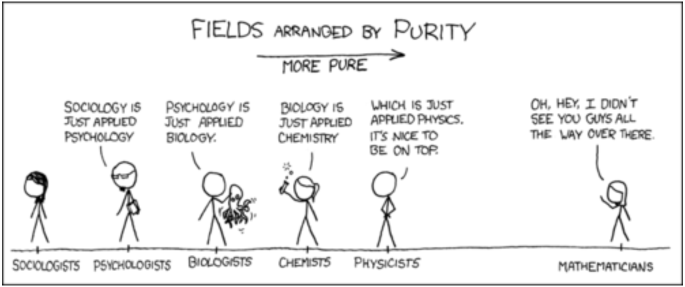
Some models have a natural hierarchical structure.



Example: study of the effectiveness of cardiac treatments. $\theta_j =$ survival proba. in hospital j. The θ_j are related. Model as sampled from some common distribution.

Hierarchical models

Some models have a natural hierarchical structure.



Example: study of the effectiveness of cardiac treatments. $\theta_j =$ survival proba. in hospital j. The θ_j are related. Model as sampled from some common distribution. The data $x_{i,j}$ (survival of patient i in hospital j) are naturally clustered. Other example : meta-analysis.

Example: Meta-analysis

To evaluate a drug for possible clinical application, a study is performed on rodents. For a particular study drawn from literature, the aim is to evaluate θ, the probability of tumor in a control population (no treatment). The data shows that 4 out of 14 rats developed a type of tumor.

Example: Meta-analysis

To evaluate a drug for possible clinical application, a study is performed on rodents. For a particular study drawn from literature, the aim is to evaluate θ, the probability of tumor in a control population (no treatment). The data shows that 4 out of 14 rats developed a type of tumor. Assume # tumors $\sim$ Bin(θ) and a conjugate prior $\theta \sim$ Beta(α, β).

Example: Meta-analysis

To evaluate a drug for possible clinical application, a study is performed on rodents. For a particular study drawn from literature, the aim is to evaluate θ, the probability of tumor in a control population (no treatment). The data shows that 4 out of 14 rats developed a type of tumor. Assume # tumors $\sim$ Bin(θ) and a conjugate prior $\theta \sim$ Beta(α, β). Posterior is

$$p(\theta|y) = Beta(\alpha + 4, \beta + 10).$$

Example: Meta-analysis

To evaluate a drug for possible clinical application, a study is performed on rodents. For a particular study drawn from literature, the aim is to evaluate θ, the probability of tumor in a control population (no treatment). The data shows that 4 out of 14 rats developed a type of tumor. Assume # tumors $\sim$ Bin(θ) and a conjugate prior $\theta \sim$ Beta(α, β). Posterior is

$$p(\theta|y) = Beta(\alpha + 4, \beta + 10).$$

Previous experiments:

0/20	0/20	0/20	0/20	0/20	0/20	0/20	0/19	0/19	0/19
0/19	0/18	0/18	0/17	1/20	1/20	1/20	1/20	1/19	1/19
1/18	1/18	2/25	2/24	2/23	2/20	2/20	2/20	2/20	2/20
2/20	1/10	5/49	2/19	5/46	3/27	2/17	7/49	7/47	3/20
3/20	2/13	9/48	10/50	4/20	4/20	4/20	4/20	4/20	4/20
4/20	10/48	4/19	4/19	4/19	5/22	11/46	12/49	5/20	5/20
6/23	5/19	6/22	6/20	6/20	6/20	16/52	15/47	15/46	9/24

Current experiment:
4/14

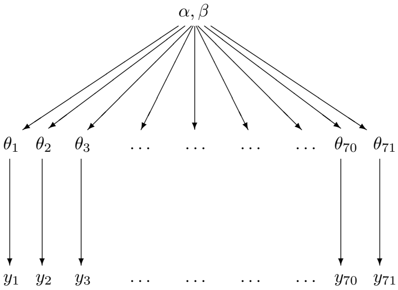
Table 5.1 *Tumor incidence in historical control groups and current group of rats, from Tarone (1982). The table displays the values of $\frac{y_j}{n_j}$: (number of rats with tumors)/(total number of rats).*

Non-Bayesian approach

Not Bayesian since not based on a full probability model.

Non-Bayesian approach

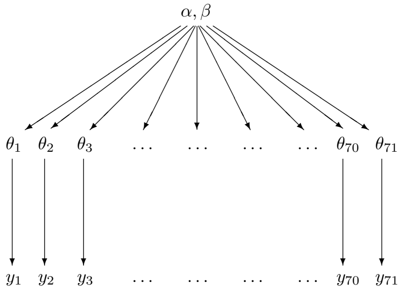
Not Bayesian since not based on a full probability model.



Observed sample mean and standard deviation of the 70 y_i/n_i are 0.136 and 0.103.

Non-Bayesian approach

Not Bayesian since not based on a full probability model.

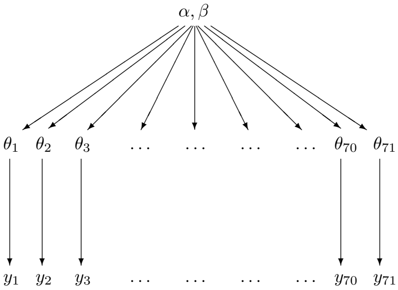


Observed sample mean and standard deviation of the 70 y_i/n_i are 0.136 and 0.103.

Pick $\hat{\alpha}, \hat{\beta}$ to match empirical mean variance of the n_j/θ_j. Get $\hat{\alpha} = 1.4, \hat{\beta} = 8.6, p(\theta|y) \sim Beta(5.4, 18.6)$

Non-Bayesian approach

Not Bayesian since not based on a full probability model.



Observed sample mean and standard deviation of the 70 y_i/n_i are 0.136 and 0.103.

Pick $\hat{\alpha}, \hat{\beta}$ to match empirical mean variance of the n_j/θ_j. Get $\hat{\alpha} = 1.4, \hat{\beta} = 8.6, p(\theta|y) \sim Beta(5.4, 18.6)$
Posterior mean is 0.223, lower than $4/14 = 0.286$. Current experiment has unusually high number of tumors.

Bayesian Hierarchical model

Setting: J exepriments, observations $y_1, \dots, y_J$ with likelihoods $L(y_j, \theta_j)$. key: specify a full probabilistic model for the θ_j.

Bayesian Hierarchical model

Setting: J exepriments, observations $y_1, \dots, y_J$ with likelihoods $L(y_j, \theta_j)$. key: specify a full probabilistic model for the θ_j. If the data is symmetric (i.e. there is no order on the experiments), then must assune that the distribution of the vector $(\theta_1, \dots, \theta_J)$ is symetric, i.e. exchangeable (invariant under relabelling by a permutation).

Bayesian Hierarchical model

Setting: J exeprimemts, observations $y_1, \ldots, y_J$ with likelihoods $L(y_j, \theta_j)$. key: specify a full probabilistic model for the θ_j. If the data is symmetric (i.e. there is no order on the experiments), then must assune that the distribution of the vector $(\theta_1, \ldots, \theta_J)$ is symetric, i.e. exchangeable (invariant under relabelling by a permutation).
Example: mixture of iid distributions

$$p(\theta) = \int \left[\prod_{j=1}^{J} \pi(\theta_j | \psi) \right] g(\psi) d\phi$$

Bayesian Hierarchical model

Setting: J exeperiments, observations $y_1, \dots, y_J$ with likelihoods $L(y_j, \theta_j)$. key: specify a full probabilistic model for the θ_j. If the data is symmetric (i.e. there is no order on the experiments), then must assune that the distribution of the vector $(\theta_1, \dots, \theta_J)$ is symetric, i.e. exchangeable (invariant under relabelling by a permutation).
Example: mixture of iid distributions

$$p(\theta) = \int \left[\prod_{j=1}^{J} \pi(\theta_j | \psi) \right] g(\psi) d\phi$$

The θ_j are drawn from a common population determined by the unknown hyperparameter ψ with *hyperprior* $g(\psi)$.

Bayesian Hierarchical model

Setting: J exepriments, observations $y_1, \dots, y_J$ with likelihoods $L(y_j, \theta_j)$. key: specify a full probabilistic model for the θ_j. If the data is symmetric (i.e. there is no order on the experiments), then must assune that the distribution of the vector $(\theta_1, \dots, \theta_J)$ is symetric, i.e. exchangeable (invariant under relabelling by a permutation).
Example: mixture of iid distributions

$$p(\theta) = \int \left[\prod_{j=1}^{J} \pi(\theta_j | \psi) \right] g(\psi) d\phi$$

The θ_j are drawn from a common population determined by the unknown hyperparameter ψ with *hyperprior* $g(\psi)$.

Remark: De Finetti's Theorem states that all exchangeable distributions are of this form in the large sample limit.

Bayesian Hierarchical model

These are models where the prior has parameters which again have a probability distribution.

- Data x have a density $f(x; \theta)$.

Bayesian Hierarchical model

These are models where the prior has parameters which again have a probability distribution.

- Data x have a density $f(x;\theta)$. (In example : $y \sim B(n,\theta)$)

Bayesian Hierarchical model

These are models where the prior has parameters which again have a probability distribution.

- Data x have a density $f(x;\theta)$. (In example : $y \sim B(n, \theta)$)
- The prior dist. of θ is $\pi(\theta;\psi)$.

Bayesian Hierarchical model

These are models where the prior has parameters which again have a probability distribution.

- Data x have a density $f(x;\theta)$. (In example : $y \sim B(n,\theta)$)
- The prior dist. of θ is $\pi(\theta;\psi)$. (In example: $\psi = (\alpha,\beta)$ and $\theta \sim Beta(\psi)$)

Bayesian Hierarchical model

These are models where the prior has parameters which again have a probability distribution.

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

Bayesian Hierarchical model

These are models where the prior has parameters which again have a probability distribution.

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

Bayesian Hierarchical model

These are models where the prior has parameters which again have a probability distribution.

- Data x have a density $f(x;\theta)$. (In example : $y \sim B(n,\theta)$)
- The prior dist. of θ is $\pi(\theta;\psi)$. (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: $p(\theta,\psi) = \pi(\theta|\psi)g(\psi)$

Bayesian Hierarchical model

These are models where the prior has parameters which again have a probability distribution.

- Data x have a density $f(x;\theta)$. (In example : $y \sim B(n,\theta)$)
- The prior dist. of θ is $\pi(\theta;\psi)$. (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: $p(\theta,\psi) = \pi(\theta|\psi)g(\psi)$
- Joint posterior: $\pi(\theta,\psi|x) \propto f(x;\theta)\pi(\theta;\psi)g(\psi)$,

Bayesian Hierarchical model

These are models where the prior has parameters which again have a probability distribution.

- Data x have a density $f(x;\theta)$. (In example : $y \sim B(n,\theta)$)
- The prior dist. of θ is $\pi(\theta;\psi)$. (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: $p(\theta,\psi) = \pi(\theta|\psi)g(\psi)$
- Joint posterior: $\pi(\theta,\psi|x) \propto f(x;\theta)\pi(\theta;\psi)g(\psi)$,
- θ prior: $\pi(\theta) = \int \pi(\theta;\psi)g(\psi)d\psi$

Bayesian Hierarchical model

These are models where the prior has parameters which again have a probability distribution.

- Data x have a density $f(x;\theta)$. (In example : $y \sim B(n,\theta)$)
- The prior dist. of θ is $\pi(\theta;\psi)$. (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: $p(\theta,\psi) = \pi(\theta|\psi)g(\psi)$
- Joint posterior: $\pi(\theta,\psi|x) \propto f(x;\theta)\pi(\theta;\psi)g(\psi)$,
- θ prior: $\pi(\theta) = \int \pi(\theta;\psi)g(\psi)d\psi$
- θ posterior $\pi(\theta|x) = \int_\Psi \pi(\theta,\psi|x)d\psi \propto f(x;\theta)\pi(\theta)$

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 $p(y|\theta) \times \pi(\theta|\psi) \times g(\psi)$.

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 $p(y|\theta) \times \pi(\theta|\psi) \times g(\psi)$.
2. Determine $\pi(\theta|\psi, y)$ (the conditional posterior density of θ given ψ for fixed observation 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 $p(y|\theta) \times \pi(\theta|\psi) \times g(\psi)$.
2. Determine $\pi(\theta|\psi, y)$ (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.

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 $p(y|\theta) \times \pi(\theta|\psi) \times g(\psi)$. Immediate
2. Determine $\pi(\theta|\psi, y)$ (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.

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 $p(y|\theta) \times \pi(\theta|\psi) \times g(\psi)$. Immediate
2. Determine $\pi(\theta|\psi, y)$ (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.

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 $p(y|\theta) \times \pi(\theta|\psi) \times g(\psi)$. Immediate
2. Determine $\pi(\theta|\psi, y)$ (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

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 $p(y|\theta) \times \pi(\theta|\psi) \times g(\psi)$. Immediate
2. Determine $\pi(\theta|\psi, y)$ (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

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

Careful about normalizing factors.

Example: Risk of tumors in rats 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.

Example: Risk of tumors in rats 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

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

Example: Risk of tumors in rats 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

$$p(\theta, \alpha, \beta | y) \propto p(\alpha, \beta)\pi(\theta|\alpha, \beta)p(y|\theta, \alpha, \beta)$$

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

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: Risk of tumors in rats cont'd

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

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

Example: Risk of tumors in rats 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 ($\rightarrow$ proper posterior?)
- recall that mean is α/β and that $\alpha+\beta$ is 'sample size'. Take log to put on a $(-\infty, \infty)$ scale and then uniform on $(\log(\alpha/\beta), \log(\alpha+\beta))$.

Example: Risk of tumors in rats 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 ($\rightarrow$ proper posterior?)
- recall that mean is α/β and that $\alpha + \beta$ is 'sample size'. Take log to put on a $(-\infty, \infty)$ scale and then uniform on $(\log(\alpha/\beta), \log(\alpha+\beta))$.$\Rightarrow$ improper posterior!
- reasonable choice of diffuse hyperprior density is uniform on $(\alpha/(\alpha+\beta), (\alpha+\beta)^{-1/2})$ which translates to $p(\alpha, \beta) \propto (\alpha+\beta)^{-5/2}$.

Example: 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.

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

Example: 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.
Qustion: 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 experiment 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

But we could also interpolate

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

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, $\theta_i \sim N(\phi, \tau^2)$.

Example: Hierarchical model for normal data

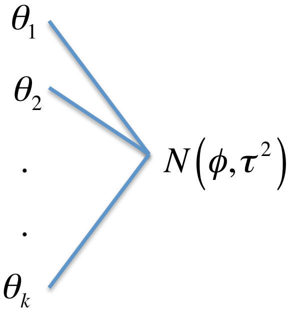
For $i = 1, 2, ..., k$ we make n_i observations $X_{i,1}, X_{i,2}, ..., X_{i,n_i}$ on population i, with $X_{ij} \sim N(\theta_i, \sigma^2)$. The θ_i are the unknown means for observations on the i'th population but σ^2 is known. Suppose the prior model for the θ_i is iid normal, $\theta_i \sim N(\phi,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

$$X_{1,1}, \ldots, X_{1,n_1} \sim N\left(\theta_1, \sigma^2\right)$$

$$X_{2,1}, \ldots, X_{2,n_2} \sim N\left(\theta_2, \sigma^2\right)$$

$$\vdots$$

$$X_{k,1}, \ldots, X_{k,n_k} \sim N\left(\theta_k, \sigma^2\right)$$



Example: Hierarchical model for normal data

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

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

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.

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

$$\pi(\theta, \psi | x) \propto f(x; \theta) \pi(\theta|\psi) g(\psi)$$

$$\propto g(\psi) \prod_{i=1}^{J} N(\theta_j | \phi, \tau^2) \prod_{j=1}^{J} N(\bar{y}_{\cdot j} | \theta_j, \sigma_j^2)$$

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

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

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

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

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.

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

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

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.

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

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

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 $\mu = \hat{\mu}$.

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 $p(\tau) \propto 1$ yields further simplification. We now want the posterior of θ given the observations y.

Example: Hierarchical model for normal data

Using $p(\tau) \propto 1$ yields further simplification. We now want the posterior of θ given the observations y.
Either

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

Example: Hierarchical model for normal data

Using $p(\tau) \propto 1$ yields further simplification. We now want the posterior of θ given the observations y.
Either

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

Example: Hierarchical model for normal data

Using $p(\tau) \propto 1$ yields further simplification. 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

Example: Hierarchical model for normal data

Using $p(\tau) \propto 1$ yields further simplification. 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

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

Example: Hierarchical model for normal data

Using $p(\tau) \propto 1$ yields further simplification. 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

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

Example: Hierarchical model for normal data

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

Integrate out wrt ϕ and τ^2 to obtain the marginal posterior distribution of θ.

Example: Hierarchical model for normal data

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

Integrate out wrt ϕ and τ^2 to obtain the marginal posterior distribution of θ. Exercise Integrating the last factor wrt ϕ gives a term proportional to

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

Example: Hierarchical model for normal data

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

Integrate out wrt ϕ and τ^2 to obtain the marginal posterior distribution of θ. Exercise Integrating the last factor wrt ϕ gives a term proportional to

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

Exercise Then the integral wrt τ gives a term proportional to

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

Example: Hierarchical model for normal data

Thus the posterior distribution of θ is

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

Example: Hierarchical model for normal data

Thus the posterior distribution of θ is

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

Let $\widehat{\theta}_j$ be the MAP estimate for θ_j (posterior mode) and put $\widehat{\theta}^* = \sum \widehat{\theta}_j / k$.

Example: Hierarchical model for normal data

Thus the posterior distribution of θ is

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

Let $\widehat{\theta}_j$ be the MAP estimate for θ_j (posterior mode) and put $\widehat{\theta}^* = \sum \widehat{\theta}_j / k$.
Exercise Differentiate $\pi(\theta|x)$ wrt θ_j and set it equal zero to get

$$\widehat{\theta}_j = (\sigma_j^{-2}\bar{x}_j + \nu\widehat{\theta}^*)/(\sigma_j^{-2} + \nu),$$

where $\nu = \left[\sum(\widehat{\theta}_i - \widehat{\theta}^*)^2/(J-2)\right]^{-1}$.

Example: Hierarchical model for normal data

Thus the posterior distribution of θ is

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

Let $\widehat{\theta}_j$ be the MAP estimate for θ_j (posterior mode) and put $\widehat{\theta}^* = \sum \widehat{\theta}_j / k$.
Exercise Differentiate $\pi(\theta|x)$ wrt θ_j and set it equal zero to get

$$\widehat{\theta}_j = (\sigma_j^{-2}\bar{x}_j + \nu\widehat{\theta}^*)/(\sigma_j^{-2} + \nu),$$

where $\nu = \left[\sum(\widehat{\theta}_i - \widehat{\theta}^*)^2/(J-2)\right]^{-1}$.
If the θ_j were unrelated then $\hat{\theta}_j = \bar{x}_j$. The model modifies the estimate by pulling it towards the mean of the estimated θ_is. Another kind of interpolation model