UPPSALA UNIVERSITET

- Hierarchical models
  - Rats example
  - Factory example
  - 8 schools example
- Exchangeability
- Computational aspects

Bayesian Statistics and Data Analysis
Lecture 7

Måns Magnusson
Department of Statistics, Uppsala University
Thanks to Aki Vehtari, Aalto University

UPPSALA UNIVERSITET

Section 1

Hierarchical models

Hierarchical model

- Example: Treatment effectiveness
  - in hospital j the survival probability is θ_j
  - observations y_{ij} tell whether patient i survived in hospital j

$$
\begin{tikzcd}
\theta_1 \arrow[d] & \theta_2 \arrow[d] & \cdots & \theta_n \arrow[d] \\
y_{i1} & y_{i2} & & y_{in}
\end{tikzcd}
$$

UPPSALA
UNIVERSITET

Hierarchical model

- Example: Treatment effectiveness
  - in hospital j the survival probability is θ_j
  - observations y_{ij} tell whether patient i survived in hospital j

$$
\begin{tikzcd}
\theta_1 \arrow[d] & \theta_2 \arrow[d] & \cdots & \theta_n \arrow[d] \\
y_{i1} & y_{i2} & & y_{in}
\end{tikzcd}
$$

  - sensible to assume that θ_j are similar

$$
\begin{tikzcd}
 & \tau \arrow[dl] \arrow[d] \arrow[drr] & & \\
\theta_1 \arrow[d] & \theta_2 \arrow[d] & \cdots & \theta_n \arrow[d] \\
y_{i1} & y_{i2} & & y_{in}
\end{tikzcd}
$$

  - natural to think that θ_j have common population distribution
  - θ_j is not directly observed and the population distribution is unknown

Hierarchical model: terms

Lvl 1: observations given parameters $p(y_{ij}|\theta_j)$

$$\begin{tikzcd}
p(\tau) & & & \tau \arrow[dll] \arrow[dl] \arrow[drr] & & & \text{hyperparameter} \\
p(\theta_j|\tau) & & \theta_1 \arrow[d] & \theta_2 \arrow[d] & \cdots & \theta_n \arrow[d] & \text{parameters} \\
p(y_{ij}|\theta_j) & & y_{i1} & y_{i2} & & y_{in} & \text{observations}
\end{tikzcd}$$

Joint posterior

$$\begin{aligned}
p(\theta, \tau|y) &\propto p(y|\theta, \tau)p(\theta, \tau) \\
&\propto p(y|\theta)p(\theta|\tau)p(\tau)
\end{aligned}$$

Hierarchical model: terms

Lvl 1: observations given parameters $p(y_{ij}|\theta_j)$

Lvl 2: parameters given hyperparameters $p(\theta_j|\tau)$

$$
\begin{tikzcd}
p(\tau) & & \tau \arrow[dl] \arrow[d] \arrow[drr] & & & \text{hyperparameter} \\
p(\theta_j|\tau) & \theta_1 \arrow[d] & \theta_2 \arrow[d] & \cdots & \theta_n \arrow[d] & \text{parameters} \\
p(y_{ij}|\theta_j) & y_{i1} & y_{i2} & & y_{in} & \text{observations}
\end{tikzcd}
$$

Joint posterior

$$
\begin{aligned}
p(\theta, \tau|y) &\propto p(y|\theta, \tau)p(\theta, \tau) \\
&\propto p(y|\theta)p(\theta|\tau)p(\tau)
\end{aligned}
$$

Comparisons

- "Separate model" (model with separate/independent effects)

$$
\begin{array}{cccc}
\theta_1 & \theta_2 & \cdots & \theta_n \\
\downarrow & \downarrow & & \downarrow \\
y_1 & y_2 & & y_n
\end{array}
$$

Comparisons

- "Separate model" (model with separate/independent effects)

$$\begin{tikzcd}
\theta_1 \arrow[d] & \theta_2 \arrow[d] & \cdots & \theta_n \arrow[d] \\
y_1 & y_2 & & y_n
\end{tikzcd}$$

- "Joint/pooled model" (model with a common effect / pooled model)

$$\begin{tikzcd}
 & \theta \arrow[dl] \arrow[d] \arrow[drr] & & \\
y_1 & y_2 & \cdots & y_n
\end{tikzcd}$$

Comparisons

- "Separate model" (model with separate/independent effects)

$$\begin{tikzcd}
\theta_1 \arrow[d] & \theta_2 \arrow[d] & \cdots & \theta_n \arrow[d] \\
y_1 & y_2 & & y_n
\end{tikzcd}$$

- "Joint/pooled model" (model with a common effect / pooled model)

$$\begin{tikzcd}
 & \theta \arrow[dl] \arrow[d] \arrow[drr] & & \\
y_1 & y_2 & \cdots & y_n
\end{tikzcd}$$

- Hierarchical model

$$\begin{tikzcd}
 & \tau \arrow[dl] \arrow[d] \arrow[drr] & & \\
\theta_1 \arrow[d] & \theta_2 \arrow[d] & \cdots & \theta_n \arrow[d] \\
y_1 & y_2 & & y_n
\end{tikzcd}$$

Predictive distributions for hiearchical models

- Two types of predictive distributions
  1. A new observation in an existing group
  2. A new observation in a new group

$$\begin{tikzcd}
& \tau \arrow[dl] \arrow[d] \arrow[drr] & & \\
\theta_1 \arrow[d] & \theta_2 \arrow[d] & \cdots & \theta_n \arrow[d] \\
y_1 & y_2 & & y_n
\end{tikzcd}$$

Hierarchical binomial model: rats

- Medicine testing
- Type F344 female rats in control group given placebo
  - count how many get endometrial stromal polyps
  - familiar binomial model example

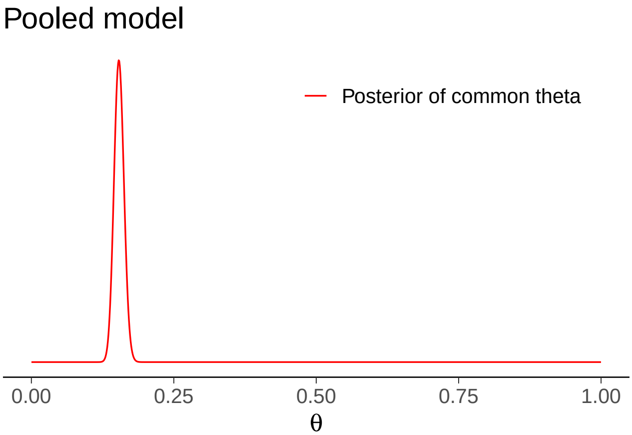
Hierarchical binomial model: rats

- Medicine testing
- Type F344 female rats in control group given placebo
  - count how many get endometrial stromal polyps
  - familiar binomial model example
- Experiment has been repeated 71 times

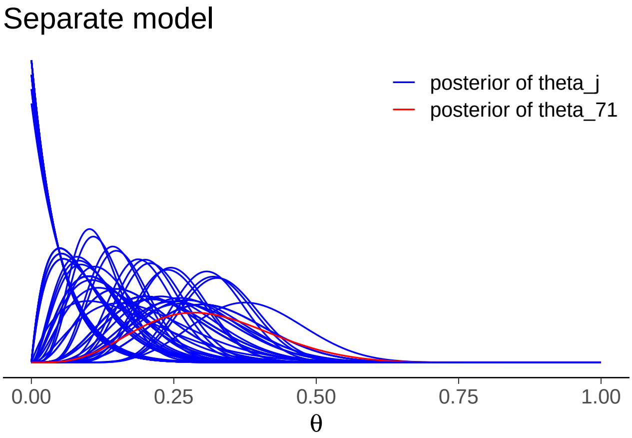
0/20	0/20	0/20	0/20	0/20	0/20	0/20	0/19	0/19	0/19
0/19	0/18	0/18	0/17	1/20	1/20	1/20	1/20	1/19	1/19
1/18	1/18	2/25	2/24	2/23	2/20	2/20	2/20	2/20	2/20
2/20	1/10	5/49	2/19	5/46	3/27	2/17	7/49	7/47	3/20
3/20	2/13	9/48	10/50	4/20	4/20	4/20	4/20	4/20	4/20
4/20	10/48	4/19	4/19	4/19	5/22	11/46	12/49	5/20	5/20
6/23	5/19	6/22	6/20	6/20	6/20	16/52	15/46	15/47	9/24
4/14									

UPPSALA UNIVERSITET

Hierarchical binomial model: rats



Hierarchical binomial model: rats

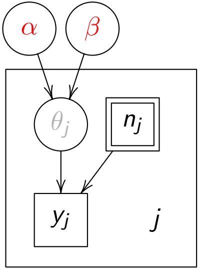


Hierarchical binomial model: rats

- Hierarchical binomial model for rats
  prior parameters α and β are unknown

$$\theta_j | \alpha, \beta \sim \text{Beta}(\theta_j | \alpha, \beta)$$

$$y_j | n_j, \theta_j \sim \text{Bin}(y_j | n_j, \theta_j)$$



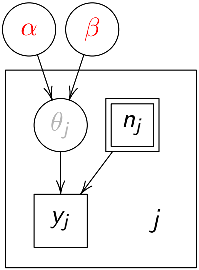
- Joint posterior $p(\theta_1, \ldots, \theta_J, \alpha, \beta | y)$
  - multiple parameters

Hierarchical binomial model: rats

- Hierarchical binomial model for rats
  prior parameters α and β are unknown

$$\theta_j|\alpha,\beta \sim \mathrm{Beta}(\theta_j|\alpha,\beta)$$

$$y_j|n_j,\theta_j \sim \mathrm{Bin}(y_j|n_j,\theta_j)$$



- Joint posterior $p(\theta_1,\ldots,\theta_J,\alpha,\beta|y)$
  - multiple parameters
  - factorize $\prod_{j=1}^{J} p(\theta_j|\alpha,\beta,y)p(\alpha,\beta|y)$

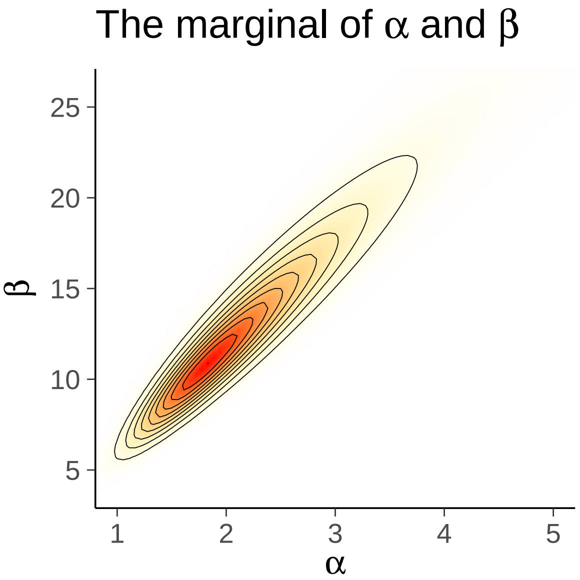
Hierarchical binomial model: rats

- Population prior Beta($\theta_j|\alpha, \beta$)
- Hyperprior $p(\alpha, \beta)$?
  - α, β both affect the location and scale
  - BDA3 (p. 110) has (vague) $p(\alpha, \beta) \propto (\alpha + \beta)^{-5/2}$

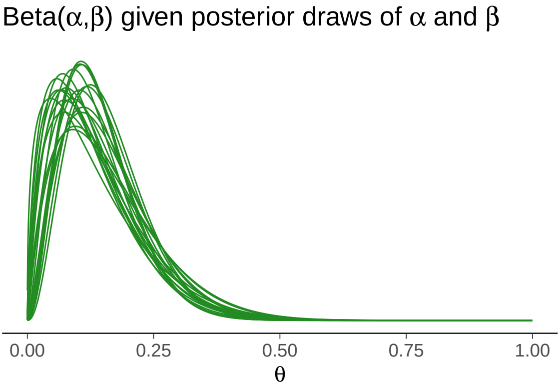
Hierarchical binomial model: rats

- Population prior Beta$(\theta_j|\alpha, \beta)$
- Hyperprior $p(\alpha, \beta)$?
  - α, β both affect the location and scale
  - BDA3 (p. 110) has (vague) $p(\alpha, \beta) \propto (\alpha + \beta)^{-5/2}$
- What type of predicitive distributions can we have?

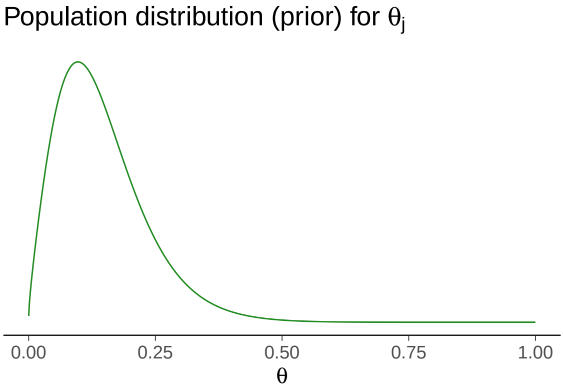
Hierarchical binomial model: rats



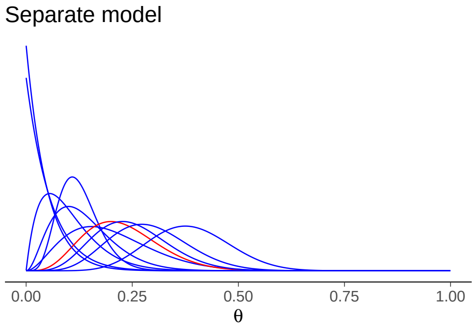
Hierarchical binomial model: rats



Hierarchical binomial model: rats

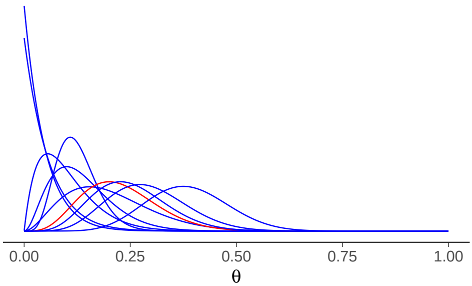


Hierarchical binomial model: rats

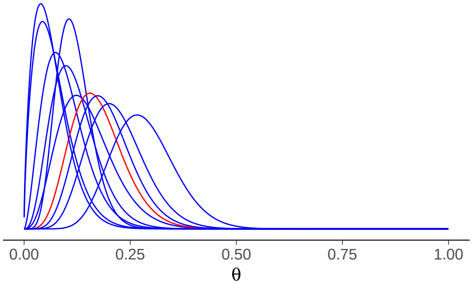


Hierarchical binomial model: rats

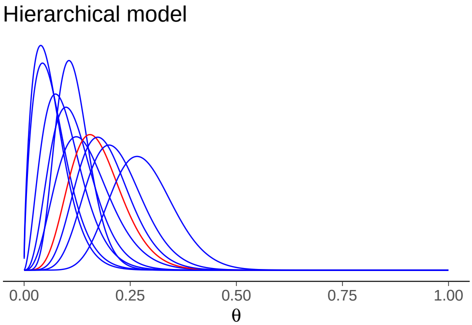
Separate model

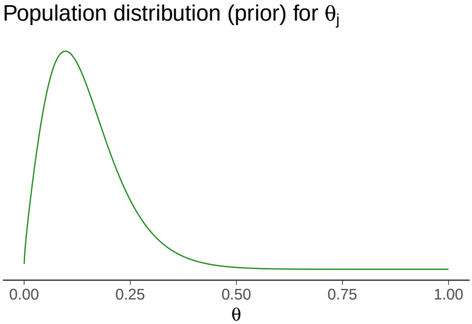


Hierarchical model



Hierarchical binomial model: rats



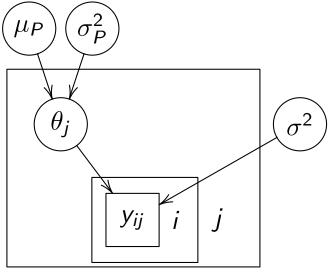


Hierarchical normal model: factory

- Factory has 6 machines which quality is evaluated
- Assume hierarchical model
  - each machine has its own (average) quality θ_j and common variance σ^2

$$\theta_j|\mu_P, \sigma_P^2 \sim \mathcal{N}(\mu_P, \sigma_P^2)$$

$$y_{ij}|\theta_j \sim \mathcal{N}(\theta_j, \sigma^2)$$

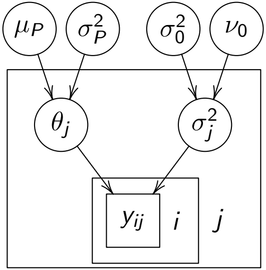


- Can be used to predict the future quality produced by each machine and quality produced by a new similar machine

Hierarchical normal model: factory

- Factory has 6 machines which quality is evaluated
- Assume hierarchical model
  - each machine has its own (average) quality θ_j and own variance σ_j^2

$$\theta_j|\mu_P, \sigma_P^2 \sim \mathcal{N}(\mu_P, \sigma_P^2)$$



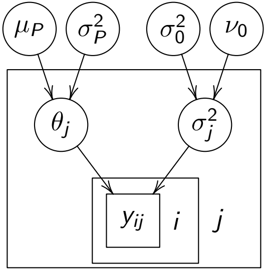
$$\sigma_j^2|\sigma_0^2, \nu_0 \sim \text{Inv-}$$

$$y_{ij}|\theta_j \sim \mathcal{N}(\theta_j, \sigma_j^2)$$

Hierarchical normal model: factory

- Factory has 6 machines which quality is evaluated
- Assume hierarchical model
  - each machine has its own (average) quality θ_j and own variance σ_j^2

$$\theta_j|\mu_P, \sigma_P^2 \sim \mathcal{N}(\mu_P, \sigma_P^2)$$



$$\sigma_j^2|\sigma_0^2, \nu_0 \sim \text{Inv-}$$

$$y_{ij}|\theta_j \sim \mathcal{N}(\theta_j, \sigma_j^2)$$

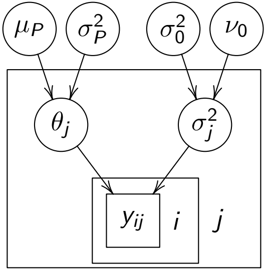
- What type of predicitive distributions can we have?

UPPSALA UNIVERSITET

Hierarchical normal model: factory

- Factory has 6 machines which quality is evaluated
- Assume hierarchical model
  - each machine has its own (average) quality θ_j and own variance σ_j^2

$$\theta_j|\mu_P, \sigma_P^2 \sim \mathcal{N}(\mu_P, \sigma_P^2)$$



$$\sigma_j^2|\sigma_0^2, \nu_0 \sim \text{Inv-}$$

$$y_{ij}|\theta_j \sim \mathcal{N}(\theta_j, \sigma_j^2)$$

- What type of predicitive distributions can we have?
- Can be used to predict the future quality produced by each machine and quality produced by a new similar machine

Hierarchical normal model: 8 schools

- Example: SAT coaching effectiveness
  - in USA commonly used Scholastic Aptitude Test (SAT) is designed so that short term practice should not improve the results significantly
  - schools have anyway coaching courses
  - test the effectiveness of the coaching courses

Hierarchical normal model: 8 schools

- Example: SAT coaching effectiveness
  - in USA commonly used Scholastic Aptitude Test (SAT) is designed so that short term practice should not improve the results significantly
  - schools have anyway coaching courses
  - test the effectiveness of the coaching courses
- SAT
  - standardized multiple choice test
  - mean about 500 and standard deviation about 100
  - most scores between 200 and 800
  - different topics, e.g., V=Verbal, M=Mathematics
  - pre-test PSAT

Hierarchical normal model: 8 schools

- Effectiveness of the SAT coaching
  - students had made pre-tests PSAT-M and PSAT-V
  - part of students were coached
  - linear regression was used to estimate the coaching effect y_j for the school j (could be denoted with $\bar{y}_{.j}$, too) and variances σ_j^2
  - y_j approximately normally distributed, with variances assumed to be known based on about 30 students per school
  - data is group means and variances (not personal results)

Hierarchical normal model: 8 schools

UPPSALA UNIVERSITET

- Effectiveness of the SAT coaching
  - students had made pre-tests PSAT-M and PSAT-V
  - part of students were coached
  - linear regression was used to estimate the coaching effect y_j for the school j (could be denoted with $\bar{y}_{.j}$, too) and variances σ_j^2
  - y_j approximately normally distributed, with variances assumed to be known based on about 30 students per school
  - data is group means and variances (not personal results)
- Data:

School	A	B	C	D	E	F	G	H
y_j	28	8	-3	7	-1	1	18	12
σ_j	15	10	16	11	9	22	20	28

Hierarchical normal model for group means

- J experiments, unknown θ_j and known σ^2

$$y_{ij}|\theta_j \sim \mathcal{N}(\theta_j, \sigma^2), \quad i = 1, \ldots, n_j; \quad j = 1, \ldots, J$$

- Group j sample mean and sample variance

$$\begin{aligned} \bar{y}_{.j} &= \frac{1}{n_j}\sum_{i=1}^{n_j} y_{ij} \\ \sigma_j^2 &= \frac{\sigma^2}{n_j} \end{aligned}$$

Hierarchical normal model for group means

- J experiments, unknown θ_j and known σ^2

$$y_{ij}|\theta_j \sim \mathcal{N}(\theta_j, \sigma^2), \quad i = 1, \dots, n_j; \quad j = 1, \dots, J$$

- Group j sample mean and sample variance

$$\bar{y}_{.j} = \frac{1}{n_j} \sum_{i=1}^{n_j} y_{ij}$$

$$\sigma_j^2 = \frac{\sigma^2}{n_j}$$

- Use model

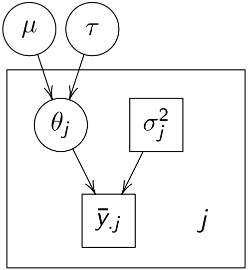
$$\bar{y}_{.j}|\theta_j \sim \mathcal{N}(\theta_j, \sigma_j^2)$$

this model can be generalized so that, σ_j^2 can be different from each other for other reasons than n_j

Hierarchical normal model for group means

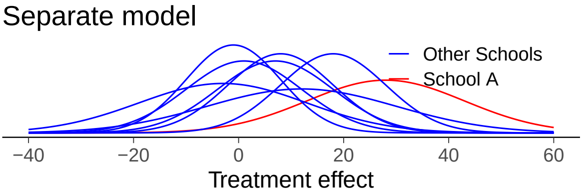
$$\theta_j|\mu,\tau \sim \mathcal{N}(\mu,\tau)$$

$$\bar{y}_{.j}|\theta_j \sim \mathcal{N}(\theta_j,\sigma_j^2)$$

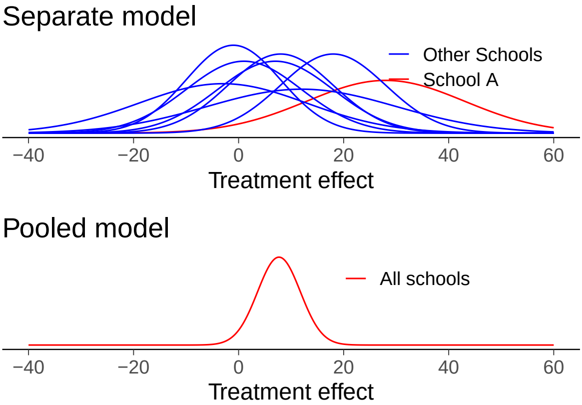


UPPSALA UNIVERSITET

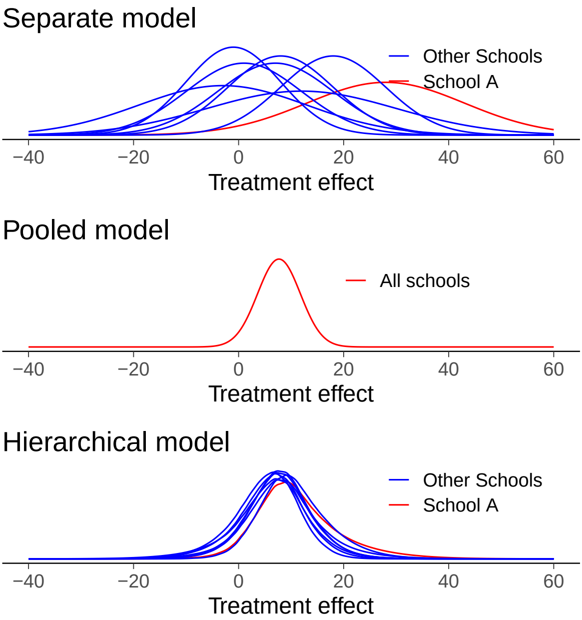
Hierarchical normal model: 8 schools



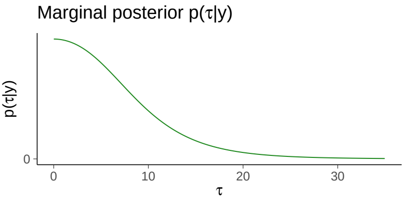
Hierarchical normal model: 8 schools



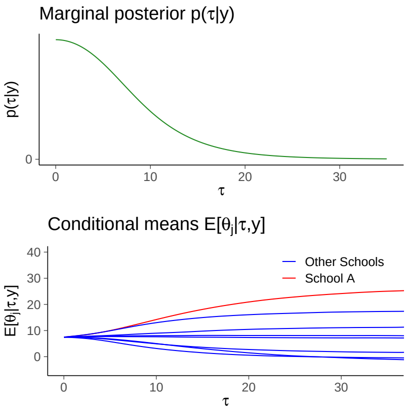
Hierarchical normal model: 8 schools



Hierarchical normal model: 8 schools

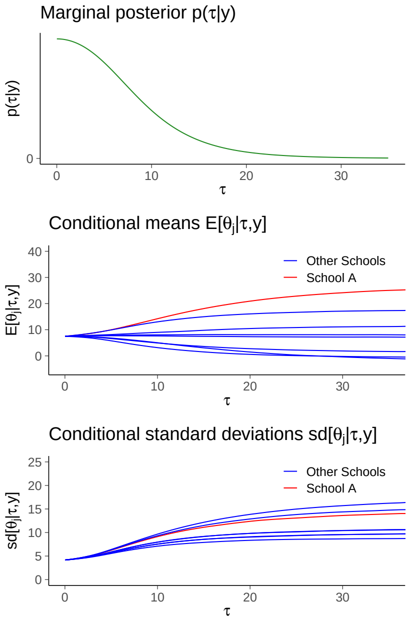


Hierarchical normal model: 8 schools



- Hierarchical models
  - Rats example
  - Factory example
  - 8 schools example
- Exchangeability
- Computational aspects

Hierarchical normal model: 8 schools



UPPSALA UNIVERSITET

- Hierarchical models
  - Rats example
  - Factory example
  - 8 schools example
- Exchangeability
- Computational aspects

Section 2

Exchangeability

UPPSALA UNIVERSITET

Exchangeability

- Justifies why we can use
  - a joint model for data
  - a joint prior for a set of parameters
- Less strict than independence (IID)
- IID implies exchangeability

Exchangeability

- *Exchangeability*
  Random variables $\theta_1, \ldots, \theta_J$ (or $y_1, \ldots, y_J$) are exchangeable if the joint distribution p is invariant to the permutation of indices
  i.e
  $$p(\theta_1, \theta_2, \theta_3) = p(\theta_{\pi(1)}, \theta_{\pi(2)}, \theta_{\pi(3)})$$
  for any permutation π of the indicies. E.g.
  $$p(\theta_1, \theta_2, \theta_3) = p(\theta_2, \theta_3, \theta_1)$$

Exchangeability

- *Exchangeability*
  Random variables $\theta_1, \ldots, \theta_J$ (or $y_1, \ldots, y_J$) are exchangeable if the joint distribution p is invariant to the permutation of indices
  i.e
  $$p(\theta_1, \theta_2, \theta_3) = p(\theta_{\pi(1)}, \theta_{\pi(2)}, \theta_{\pi(3)})$$
  for any permutation π of the indicies. E.g.
  $$p(\theta_1, \theta_2, \theta_3) = p(\theta_2, \theta_3, \theta_1)$$
  Can we come up with a situation where this doesn't hold?

Exchangeability

- *Exchangeability*
  Random variables $\theta_1, \dots, \theta_J$ (or $y_1, \dots, y_J$) are exchangeable if the joint distribution p is invariant to the permutation of indices
  i.e
  $$p(\theta_1, \theta_2, \theta_3) = p(\theta_{\pi(1)}, \theta_{\pi(2)}, \theta_{\pi(3)})$$
  for any permutation π of the indicies. E.g.
  $$p(\theta_1, \theta_2, \theta_3) = p(\theta_2, \theta_3, \theta_1)$$
  Can we come up with a situation where this doesn't hold?
  E.g. when we have a trend over indicies (like in time series data)

UPPSALA UNIVERSITET

Exchangeability

- *Exchangeability*
  Random variables $\theta_1, \ldots, \theta_J$ (or $y_1, \ldots, y_J$) are exchangeable if the joint distribution p is invariant to the permutation of indices
  i.e
  $$p(\theta_1, \theta_2, \theta_3) = p(\theta_{\pi(1)}, \theta_{\pi(2)}, \theta_{\pi(3)})$$
  for any permutation π of the indicies. E.g.
  $$p(\theta_1, \theta_2, \theta_3) = p(\theta_2, \theta_3, \theta_1)$$
  Can we come up with a situation where this doesn't hold?
  E.g. when we have a trend over indicies (like in time series data)
- Exchangeability implies symmetry:
  If there is no information which can be used *a priori* to separate θ_j form each other, we can assume exchangeability. ("Ignorance implies exchangeability")

Exchangeability: DeFinettis theorem

Let $(x_n)_{n=1}^{\infty}$ to be an infinite sequence of exchangeable random variables. De Finetti's theorem then says that there is some random variable θ so that x_j are conditionally independent given θ, and joint density for $x_1, \ldots, x_J$ can be written in the *iid mixture* form

$$p(x_1, \ldots, x_J) = \int \left[\prod_{j=1}^{J} p(x_j|\theta) \right] p(\theta) d\theta$$

- Exchangeability: the data behave as if they were i.i.d., once we account for some unknown parameter θ

Example

- Suppose we're observing coin flips, but we're not sure about the bias θ
  - We think the flips are 'the same kind of thing' (no reason to treat flip 1 differently from flip 2): Exchangeability
  - But we might not believe they are strictly independent because were uncertain about the bias: they share information through the unknown θ

UPPSALA UNIVERSITET

Example

- Suppose we're observing coin flips, but we're not sure about the bias θ
  - We think the flips are 'the same kind of thing' (no reason to treat flip 1 differently from flip 2): Exchangeability
  - But we might not believe they are strictly independent because were uncertain about the bias: they share information through the unknown θ
- Given θ the flips are i.i.d.
- Without conditioning on θ, the flips are only exchangeable, not iid.

UPPSALA UNIVERSITET

Exchangeability

- Exchangeability does not mean that the results of the experiments could not be different
  - e.g. if we know that the experiments have been in two different laboratories, and we know that the other laboratory has better conditions for the rats, but we do not know which experiments have been made in which laboratory
  - a priori experiments are exchangeable
  - model could have unknown parameter for the laboratory with a conditional prior for rats assumed to come form the same place (clustering model)

UPPSALA UNIVERSITET

Exchangeability and additional information

- Example: bioassay
  - y_i number of dead animals are not exchangeable alone

UPPSALA UNIVERSITET

Exchangeability and additional information

- Example: bioassay
  - y_i number of dead animals are not exchangeable alone
  - x_i dose is additional information

Exchangeability and additional information

- Example: bioassay
  - y_i number of dead animals are not exchangeable alone
  - x_i dose is additional information
  - (x_i, y_i) exchangeable and logistic regression was used

$$p(\alpha, \beta | y, n, x) \propto \prod_{i=1}^{n} p(y_i | \alpha, \beta, n_i, x_i) p(\alpha, \beta)$$

Hierarchical exchangeability

- Example: hierarchical rats example
  - all rats not exchangeable

Hierarchical exchangeability

- Example: hierarchical rats example
  - all rats not exchangeable
  - in a single laboratory rats exchangeable

Hierarchical exchangeability

- Example: hierarchical rats example
  - all rats not exchangeable
  - in a single laboratory rats exchangeable
  - laboratories exchangeable

Hierarchical exchangeability

- Example: hierarchical rats example
  - all rats not exchangeable
  - in a single laboratory rats exchangeable
  - laboratories exchangeable

UPPSALA UNIVERSITET

Hierarchical exchangeability

- Example: hierarchical rats example
  - all rats not exchangeable
  - in a single laboratory rats exchangeable
  - laboratories exchangeable
    $\rightarrow$ hierarchical model can be used

Partial or conditional exchangeability

- Conditional exchangeability
  - if y_i is connected to an additional information x_i, so that y_i are not exchangeable, but (y_i, x_i) exchangeable use joint model or conditional model $(y_i|x_i)$.

Partial or conditional exchangeability

- Conditional exchangeability
  - if y_i is connected to an additional information x_i, so that y_i are not exchangeable, but (y_i, x_i) exchangeable use joint model or conditional model $(y_i | x_i)$.
- Partial exchangeability
  - if the observations can be grouped (a priori), then we can use a hierarchical model

Section 3

Computational aspects

The funnel posterior

- Hiearchical models
  - Group-level or global parameters, e.g.

$$\tau \sim p(\tau)$$

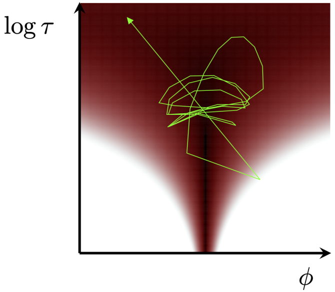
  - Local or individual-level parameters

$$\theta_i \sim \mathcal{N}(0, \tau)$$

- Creates a "funnel-like" posterior geometry:
- Comes from the variance in the different layers:
  - When τ is small, the θ_i's are concentrated around 0
  - When τ is large, the θ_i's are widely dispersed

Why problematic?

1. Pathological geometry: difficult to explore efficiently
2. Divergences: HMC will risk divergencies
   (also a good diagnostic)



Betancourt (2020)

demo

UPPSALA UNIVERSITET

- Hierarchical models
  - Rats example
  - Factory example
  - 8 schools example
- Exchangeability
- Computational aspects

Handling the funnel

1. Reduce step size (`adapt_delta` closer to 1)
2. Reparametrize using non-centered parametrization
   2.1 Centered parametrization

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

Handling the funnel

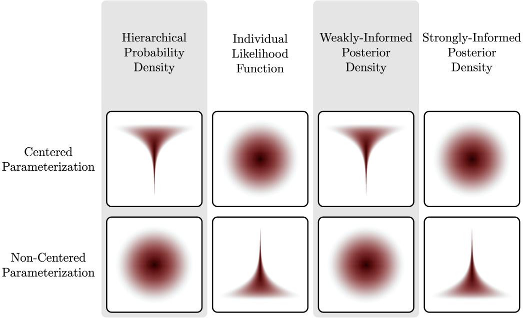
1. Reduce step size (`adapt_delta` closer to 1)
2. Reparametrize using non-centered parametrization
   2.1 Centered parametrization

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

   2.2 Non-centered parametrization

$$\eta_i \sim N(0, 1)$$

$$\theta_i = \mu + \tau\eta_i$$



Betancourt (2020)