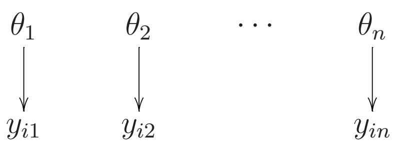
Hierarchical models (chapter 5)

- Introduction to hierarchical models
  - sometimes called multilevel model
- Exchangeability

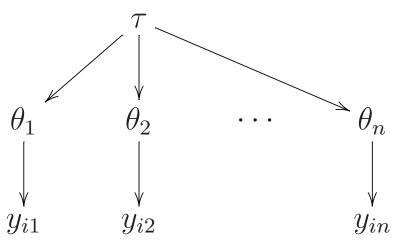
Slide 1

Hierarchical model

- Example: heart surgery in hospitals
  - in hospital j survival probability θ_j
  - observations y_{ij}, i.e. whether patient i survived in hospital j



  - natural to assume that θ_j may be different but similar



Slide 2

Slide 3

Hierarchical model: risk of tumor in rats

- Example: risk of tumor in rats
  - drugs tested on rodents before clinical trial
  - estimate the probability of tumor θ in a population of type 'F344' female laboratory rats given a zero dose (control group)
  - data: 4/14 rats developed endometrial stromal polyps
  - assume binomial and conjugate prior
  - prior?

Slide 4

Hierarchical model: risk of tumor in rats

- Previous experiments $y_1, \ldots, y_{70}$

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

- Current experiment y_{71} : 4/14
- Previously binomial $p(y_j|\theta)$, where θ common to all experiment
- Now $p(y_j|\theta_j)$, ie. every experiment has different θ_j
  - the probability of tumor θ_j vary because of differences in rats and experimental conditions

Hierarchical model

- How to take into account, that $\theta_1, \ldots, \theta_{71}$ likely similar

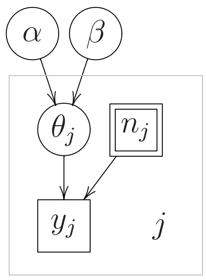
$\longrightarrow$ common population prior

- Solution is a hierarchical model

Slide 5

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

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

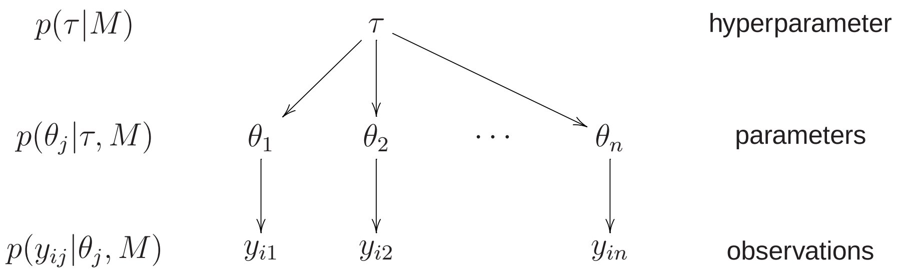


- Joint posterior $p(\theta_1, \ldots, \theta_J, \alpha, \beta|y)$
  - multiparameter model
  - factored $\prod_{j=1}^{J} p(\theta_j|\alpha,\beta,y)p(\alpha,\beta|y)$

Hierarchical model

Slide 6

- Hierarchical model:
  Level 1: observations given parameters $p(y_{ij}|\theta_j, M)$
  Level 2: parameters given hyperparameters $p(\theta_j|\tau, M)$



- Joint posterior

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

Hierarchical model: risk of tumor in rats

- Population prior $\mathrm{Beta}(\theta_j|\alpha,\beta)$
- Hyperprior $p(\alpha,\beta)$?
  - In Beta-distribution α,β both have effect on location and scale
  - Gelman et al propose prior $p(\alpha,\beta) \propto (\alpha+\beta)^{-5/2}$
    - diffuse prior on both location and scale (see p. 128)

Slide 7

- Esim6_1.m
  - hierarchical model assumes, that θ_j are similar, but not same

Hierarchical model

- Predictive distribution for a future observation $\tilde{y}$ given θ_j for current j
  - e.g. a new patient in hospital j
- Predictive distribution for a future observation $\tilde{y}$ given new θ_j for new j i.e. $\tilde{\theta}$
  - e.g. a new patient in a new hospital

Slide 8

Hierarchical model - computation

- Easy to sample from the factored distribution
  1. sample $\tilde{\phi}$ from the marginal $p(\phi|y)$
  2. sample $\tilde{\theta}$ from the conditional $p(\theta|\tilde{\phi}, y)$
  3. if needed sample $\tilde{y}$ from the predictive distribution $p(y|\tilde{\theta})$
  - repeat L times

Slide 9

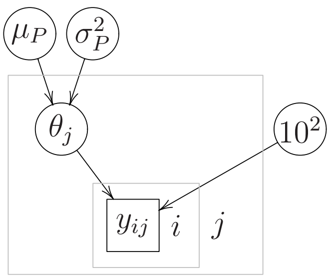
Hierarchical normal model - IQ-example

- Previously
  - population $\theta_j \sim \mathrm{N}(100, 15^2)$ and observation $y_{ij}|\theta_j \sim \mathrm{N}(\theta_j, 10^2)$
- Using hierarchical model
  - population distribution may be unknown

Slide 10

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

$$y_{ij}|\theta_j \sim \mathrm{N}(\theta_j, 10^2)$$



- Making IQ-test for several persons and using hierarchical model it is possible learn about population distribution, which then works as a prior for individual θ_j
- Measurement variance can be assumed unknown, too

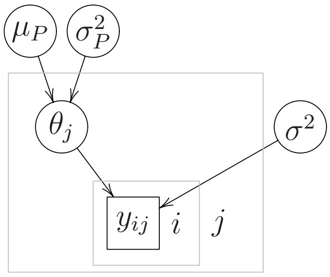
Slide 11

Hierarchical model: example

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

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

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



- Possible to predict future quality for each machine and for a new machine
- Gibbs-sampling exercise (next week)

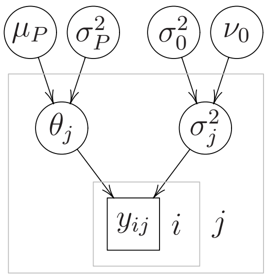
Slide 12

Hierarchical model: example

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

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

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



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

- Possible to predict future quality for each machine and for a new machine
- Gibbs-sampling exercise extra points

Hierarchical normal model - SAT-example

- Example: analyze the effects special coaching programs (ex 5.1*)
  - In USA students tested with SAT (*Scholastic Aptitude Test*), which has been designed so that short term training should not improve score
  - some schools still have short-term coaching programs
  - analyze whether coaching has any effect
- SAT
  - standardized multiple choice test
  - mean about 500 and deviation about 100
  - scores can vary between 200 and 800
  - different subjects like V=Verbal, M=Mathematics
  - preliminary test= PSAT

Slide 13

Hierarchical normal model - SAT-example

- Analyze the effect of coaching
  - students have taken PSAT-M and PSAT-V
  - part of the students were coached
  - linear regression estimates the coaching effect y_j (can be written also as $\bar{y}_{.j}$) and variances σ_j^2
  - y_j approximately normally distributed with approximately known variances based on results of about 30 students per school
  - note! data is group means and variances (not results of single students)
- 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

  - 8 points corresponds to about one correct answer

Slide 14

SAT example

- J schools, unknown θ_j and known σ^2

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

- Summarize group j with mean and variance

Slide 15

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

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

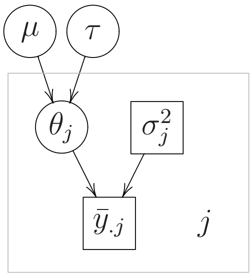
- Use model

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

Hierarchical normal model fro group means

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

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



Slide 16

Model for means

- Model

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

  - can be used for other data, where averages $\bar{y}_{.j}$ are assumed to be nearly normally distributed, even data y_{ij} are not

Slide 17

SAT example - prior

- Semiconjugate prior

$$p(\theta_1, \ldots, \theta_J|\mu, \tau) = \prod_{j=1}^{J} \mathrm{N}(\theta_j|\mu, \tau^2)$$

  - if $\tau \longrightarrow \infty$ then (*separate model*)
  - if $\tau \longrightarrow 0$, then (*pooled model*), i.e. $\theta_j = \mu$ and $\bar{y}_{.j}|\mu \sim \mathrm{N}(\mu, \sigma_j^2)$

Slide 18

SAT example - hyperprior

- Model

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

- Semi-conjugate prior

$$p(\theta_1, \ldots, \theta_J|\mu, \tau) = \prod_{j=1}^{J} \mathrm{N}(\theta_j|\mu, \tau^2)$$

Slide 19

- Hyperpior

$$p(\mu, \tau) = p(\mu|\tau)p(\tau) \propto p(\tau)$$

  - uniform prior for μ ok
  - prior for τhas to selected more carefully
  - $p(\tau) \propto 1/\tau$ would produce improper prior
  - if $J > 4$, $p(\tau) \propto 1$ reasonable uninformative prior
  - if $J \leq 4$ half-Cauchy useful (Gelman, 2005)

Hierarchical normal model – factored computation

- Factorize joint posterior

$$p(\theta, \mu, \tau|y) \propto p(\theta|\mu, \tau, y)p(\mu, \tau|y)$$

- Conditional posterior for θ_j

$$\theta_j|\mu, \tau, y \sim \mathrm{N}(\hat{\theta}_j, V_j)$$

Slide 20

where $\hat{\theta}_j$ and V_j are sane as for J independent normal distribution given informative conjugate prior

  - ie. precision weighted average of data and prior

Hierarchical normal model – factored computation

- Marginal posterior for hyperparameters

$$p(\mu, \tau|y) \propto p(\mu, \tau) \prod_{j=1}^{J} \mathrm{N}(\bar{y}_{.j}|\mu, \sigma_j^2 + \tau^2)$$

- Could be used directly (eg. with 2-dimensional grid sampling), but can be factorized

$$p(\mu, \tau|y) = p(\mu|\tau, y)p(\tau|y)$$

Slide 21

where

$$p(\mu|\tau, y) = N(\hat{\mu}, V_\mu)$$

where $\hat{\mu}$ is precision weighted mean of $\bar{y}_{.j}$ and V_μ is overall precision

- Marginal

$$p(\tau|y) = \frac{p(\mu, \tau|y)}{p(\mu|\tau, y)}$$

is not in closed form, but since unidimensional, easy to sample eg. with inverse-cdf

SAT example - computation

- Factored sampling

$$p(\theta, \mu, \tau|y) \propto p(\tau|y)p(\mu|\tau, y)p(\theta|\mu, \tau, y)$$

- Ex 5.1*
  - see "Computation" s. 137
- Esim6_2.m

Slide 22

Meta-analysis

- Meta-analysis combines and analyzes several experiments on same subjects
  - eg. in medical science several smaller experiments made in different countries
  - meta-analysis combines published results to combine information and reduce uncertainty
  - meta-analysis handled with hierarchical model
- p. 145 in book

Slide 23

Exchangeability

- Justifies why we can use
  - common model for data
  - common prior for parameters
- Less strict assumption than independency
- "Ignorance implies exchangeability"

Slide 24

Exchangeability

- Set of experiments $j = 1, \dots, J$
- Experiment j with observations y_j, parameter θ_j and model $p(y_j|\theta_j)$
- Some of the parameters can be common to all experiments
  - eg. in hierarchical normal model may be $\theta_j = (\mu_j, \sigma^2)$, assuming same variance in different experiments

Slide 25

Exchangeability

- Two ways to define
  1. If no other information – other than the data y – is available to distinguish any of the θ_j from any of the others, and no ordering or grouping of the parameters can be made, one must assume symmetry among the parameters in their prior distribution
     - this symmetry is represented probabilistically by exchangeability
  2. Parameters $\theta_1, \dots, \theta_J$ are exchangeable in their joint distribution if $p(\theta_1, \dots, \theta_J)$ is invariant to permutations of the indexes $(1, \dots, J)$

Slide 26

Exchangeability

- Exchangeability does not imply that results can not be different
  - eg. if we know that experiments have been made in two different labs with different conditions, but we don't know which experiments were made in which lab
  - a priori experiments still exchangeable
  - model might have unknown parameter telling in which lab experiment was made, and then conditionally common prior for experiments made in one lab (clustering model)

Slide 27

Exchangeability

- Simplest form of exchangeability (but not the only one) for the parameters θ is iid

$$p(\theta|\phi) = \prod_{j=1}^{J} p(\theta_j|\phi)$$

- Often ϕ unknown and we want to compute θ's marginal distribution

Slide 28

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

- This form is a mixture of iid distributions
- de Finetti's theorem states that in the limit $J \rightarrow \infty$, any suitable well-behaved exchangeable distribution on $(\theta_1, \ldots, \theta_J)$ can be written in this form
  - formally does not hold for finite J

Exchangeability vs. independence

- Example: Six sided die with probabilities $\theta_1, \ldots, \theta_6$
  - without any other knowledge $\theta_1, \ldots, \theta_6$ exchangeable
  - due to restriction $\sum_{j=1}^{6} \theta_j$ not independent and cannot be modeled as a mixture of iid distributions

Slide 29

Exchangeability

1) box has 1 black and 1 white ball, first pick one y_1, put it back , mix and pick second ball y_2
   - are observations y_1 and y_2 exchangeable?
   - are observations y_1 and y_2 independent?
2) box has 1 black and 1 white ball, first pick one ball y_1, do not put it back, and pick a second ball y_2
   - are observations y_1 and y_2 exchangeable?
   - are observations y_1 and y_2 independent?
3) box has 10000 black and 10000 white balls, first pick one ball y_1, do not put it back, and pick a second ball y_2
   - are observations y_1 and y_2 exchangeable?
   - are observations y_1 and y_2 independent?
   - can we proceed as if observations were independent?

Slide 30

Exchangeability

4) box has a few (n known) black and white balls (proportion unknown), first pick one ball y_1, put it back, mix and pick a second ball y_2
   - are observations y_1 and y_2 exchangeable?
   - are observations y_1 and y_2 independent?
   - can we proceed as if observations were independent?
5) box has a few (n known) black and white balls (proportion unknown), first pick one ball y_1, do not put it back, and pick a second ball y_2
   - are observations y_1 and y_2 exchangeable?
   - are observations y_1 and y_2 independent?
   - can we proceed as if observations were independent?

Slide 31

Exchangeability

6) box has many (n known or unknown) black and white balls (proportion unknown), first pick one ball y_1, do not put it back, and pick a second ball y_2
   - are observations y_1 and y_2 exchangeable?
   - are observations y_1 and y_2 independent?
   - can we proceed as if observations were independent?

Slide 32

Exchangeability

- Example: divorce rates per 1000 residents in 8 USA states in 1981
  - without other knowledge $y_1, \ldots, y_8$ are exchangeable
- Divorce rates of seven first states are $5.6, 6.6, 7.8, 5.6, 7.0, 7.2, 5.4$
  - $y_1, \ldots, y_8$ are exchangeable
- Alternatively known, that 8 states are Arizona, Colorado, Idaho, Montana, Nevada, New Mexico, Utah, Wyoming, but order is unknown

Slide 33

  - before seeing the data $y_1, \ldots, y_8$ still exchangeable, but prior might take into account that, there are lot of Mormons in Utah and it is easy to get divorce in Nevada; prior could be multimodal
- Alternatively known, that y_8 is Nevada
  - even before seeing the data, $y_1, \ldots, y_8$ not anymore exchangeable, because there is information which makes y_8 different from others
  - prior might be that $p(y_8 > \max(y_1, \ldots, y_7))$ is large
  - Nevada had actually 13.9 divorces per 1000 residents

Exchangeability and additional information

- Example: if divorce rate in previous year x_j in each state j were known
  - y_j are not exchangeable
  - (x_j, y_j) are exchangeable
  - generally exchangeability can achieved by conditioning on additional information

Slide 34

$$p(\theta_1, \ldots, \theta_J | x_1, \ldots, x_J) = \int \left[\prod_{j=1}^{J} p(\theta_j | \phi, x_j) \right] p(\phi | x_1, \ldots, x_J) d\phi$$

  - x_j is called *covariate*, which implies that its value variates with y_j
- This way exchangeability is general-purpose approach, because additional information can be included in x and y

Exchangeability and additional information

- Example: bioassay
  - x_i dose
  - y_i number of animals died
  - (x_i, y_i) pair is exchangeable and conditional model was used

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

Slide 35

Exchangeability and conditional modeling (s. 354)

- Joint model (x_i, y_i)

$$p(x, y | \varphi, \theta) = p(x | \varphi) p(y | x, \theta)$$

- Assume φ and θ a priori independent i.e. $p(\varphi, \theta) = p(\varphi)p(\theta)$, and thus

$$p(\varphi, \theta | x, y) = p(\varphi | x) p(\theta | x, y)$$

Slide 36

- We can examine just the term $p(\theta | x, y)$

$$p(\theta | x, y) \propto p(y | x, \theta) p(\theta)$$

- if x chosen e.g. in design of experiments, $p(x)$ does not exist or is known and does not have parameters

Hierarchical exchangeability

- Example: heart surgery
  - all patients are not exchangeable with each other
  - in single hospital patients are exchangeable (given no other information)
  - hospitals are exchangeable (given no other information)

  $\longrightarrow$ hierarchical model

Slide 37

Partial or conditional exchangeability

- Often observations not fully exchangeable
- Partial exchangeability
  - if observations can be grouped $\longrightarrow$ hierarchical model, in which groups are exchangeable and observations inside groups are exchangeable
- Conditional exchangeability
  - if y_i has related information x_i, which makes y_i not exchangeable, but (y_i, x_i) is exchangeable possible to make a joint or conditional model $(y_i|x_i)$.

Slide 38

Exchangeability

- Observations $y_1, \ldots, y_n$ are exchangeable in their joint distribution if $p(y_1, \ldots, y_n)$ is invariant to permutation of indexes $(1, \ldots, n)$
- Parameters $\theta_1, \ldots, \theta_J$ are exchangeable in their joint distribution if $p(\theta_1, \ldots, \theta_J)$ is invariant to permutation of indexes $(1, \ldots, J)$
- Simplest form of the exchangeability (not only form) is independent samples

Slide 39

$$p(y|\theta) = \prod_{i=1}^{n} p(y_i|\theta_j) \qquad \text{or} \qquad p(\theta|\phi) = \prod_{j=1}^{J} p(\theta_j|\phi)$$