§4-5 Hierarchical & Graphical Models

Outline

1. Non-hierarchical models
2. Hierarchical models (hierarchical priors)
3. Graphical models (DAG: Directed Acyclic Graph)
4. Using DAGs for hierarchical models
5. Summary

1. Non-hierarchical models

Example: Drug efficacy

Data:

$$\begin{aligned} y &= \text{15 successes from} \\ n &= \text{20 independent trials} \end{aligned}$$

Likelihood:

$$Y \mid \theta \sim \text{Binomial}(n, \theta) \ ,$$

where θ is *true* success rate (ie probability of success)

Prior:

$$\theta \sim \text{Beta}(9.2, 13.8)$$

Posterior:

$$\begin{aligned} p(\theta \mid y) &\propto p(y \mid \theta)\, p(\theta) \\ \theta \mid y &\sim \text{Beta}(24.2, 18.8) \end{aligned}$$

Example: Hospital death rates

Now suppose we observe N sets of binomial data, for example: N=12 hospitals performing cardiac surgery in babies

Number of failures (deaths) per hospital:

Hospital i	1	2	3	....	10	11	12
No. of ops. n_i	15	148	10	....	97	256	360
No. of deaths y_i	0	18	1	....	8	29	24

How would you model these data?

Assume that, given 'true' death rate θ_i (ie probability of death) in hospital i, operation outcomes within hospital i are independent.

$$Y_i \mid \theta_i \sim \text{Binomial}(n_i, \theta_i) \ \ (i = 1, \ldots, 12)$$

Using a common death rate θ

Assume true death rate in each hospital is the same (ie $\theta_i = \theta, \forall i$).

$$Y_i \mid \theta \sim \text{Binomial}(n_i, \theta) \;\; (i = 1, \ldots, 12)$$

Then, likelihood is

$$p(\mathbf{y} \mid \theta) \;\; = \;\; \prod_{i=1}^{12} p(y_i \mid \theta)$$

$$\propto \prod_{i=1}^{12} \theta^{y_i}(1-\theta)^{n_i - y_i} \;\; = \;\; \theta^{\sum y_i}(1-\theta)^{(\sum n_i - \sum y_i)}$$

This is equivalent to observing a single hospital with $\sum_i y_i$ deaths in $\sum_i n_i$ operations.

Assume Beta prior for θ with α, β fixed:

$$\theta \mid \alpha, \beta \sim \text{Beta}(\alpha, \beta)$$

Then the posterior for θ is

$$\theta \mid \mathbf{y} \;\; \sim \;\; \text{Beta}\left(\sum_{i=1}^{12} y_i + \alpha, \sum_{i=1}^{12} (n_i - y_i) + \beta\right)$$

But is it reasonable to assume a *common* probability θ of death for every hospital?

Using different death rates θ_i

In each hospital i (with 'true' death rate θ_i),

$$\begin{aligned} Y_i \mid \theta_i &\sim \text{Binomial}(n_i, \theta_i) \\ \theta_i \mid \alpha, \beta &\sim \text{Beta}(\alpha, \beta) \end{aligned}$$

- θ_i's are random sample from a common *population distribution*: Beta(α, β)
- So, hospital 'true' death rates are assumed to be **similar** but not identical.
  Is this reasonable?
  Suppose the only information you have is that 3 hospitals have 'true' death rates 5%, 4% and 9% respectively. Guess the death rate of a 4th hospital

How would you specify values for α and β?
How would you justify the values of α and β?

Empirical Bayes approach

Hospital i	1	2	3	....	10	11	12
No. of ops. n_i	15	148	10	....	97	256	360
No. of deaths y_i	0	18	1	....	8	29	24

1. Calculate observed death rates $\frac{y_i}{n_i}$
2. Calculate the mean and variance of these 12 values $\frac{y_i}{n_i}$
3. Find $\widehat{\alpha}$, $\widehat{\beta}$ such that Beta$(\widehat{\alpha}, \widehat{\beta})$ distribution has this mean and variance.
4. Use $\theta_i \sim$ Beta$(\widehat{\alpha}, \widehat{\beta})$ as a prior to obtain posterior $\theta_i \mid y_i$

Disadvantages of this approach are:

- We are using the data twice: once to estimate the prior; again in the likelihood. $\Rightarrow$ over-estimated precision of our inference
- Using a point estimate for α and β (and treating them as fixed) ignores some uncertainty about the population distribution of the θ_i's

2. Hierarchical models

Fundamental idea of Bayesian inference is to assume a probability distribution for uncertainty about any unknown quantities.

So, treat α and β as unknown and independent, and assign prior distributions to them independently, e.g.

$$\begin{aligned} \alpha &\sim \text{Exponential}(0.01) \\ \beta &\sim \text{Exponential}(0.01) \end{aligned}$$

Now, the unknown parameters are $(\alpha, \beta, \boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{12})$. Since $\theta_i \sim \text{Beta}(\alpha, \beta)$ independently for each i given α and β, the joint prior distribution for the entire set of parameters is

$$p(\boldsymbol{\theta}, \alpha, \beta) = \left\{ \prod_{i=1}^{N} p(\theta_i \mid \alpha, \beta) \right\} p(\alpha)\, p(\beta)$$

Bayes Theorem gives us the joint posterior distribution of $(\alpha, \beta, \boldsymbol{\theta})$:

$$p(\boldsymbol{\theta}, \alpha, \beta \mid \mathbf{y}) \;\propto\; \left\{ \prod_{i=1}^{N} p(\theta_i \mid \alpha, \beta) \right\} p(\alpha)\, p(\beta) \times \left\{ \prod_{i=1}^{N} p(y_i \mid \theta_i) \right\}$$

Advantages of this approach:

- The posterior distribution for each θ_i
  - *'borrows strength'* from the likelihood contributions of *all* hospitals, via their influence on the estimate of the unknown population parameters α, β
  - reflects our full uncertainty about the true values of α and β
- This latter is also useful if we are interested in α and β themselves
  (e.g. $\alpha/(\alpha+\beta)$ is mean death rate over population of hospitals)

Such models are also called *Random effect* or *Multilevel* models.

Example: Hospital death rates

In the 12 hospitals, there were a total of 2073 operations including 159 deaths; ie, the overall death rate is $159/2073 = 0.077$.

We fitted the following models:

1. MLE (non-Bayesian): y_i/n_i
2. Non-hierarchical Bayesian

$$\begin{aligned} Y_i \mid \theta_i &\sim \text{Binomial}(n_i, \theta_i) \\ \theta_i \mid \alpha, \beta &\sim \text{Beta}(\alpha = 1, \beta = 1) \end{aligned}$$

   The posterior distribution of θ_i for the non-hierarchical model is Beta(y_i+1, n_i-y_i+1). So, the posterior mean of θ_i is $E[\theta_i|\mathbf{y}] = \frac{y_i+1}{n_i+2}$.

3. Hierarchical Bayesian

$$\begin{aligned} Y_i \mid \theta_i &\sim \text{Binomial}(n_i, \theta_i) \\ \theta_i \mid \alpha, \beta &\sim \text{Beta}(\alpha, \beta) \\ \alpha &\sim \text{Exponential}(0.01) \\ \beta &\sim \text{Exponential}(0.01) \end{aligned}$$

   The hierarchical model was fitted by using WinBUGS.

Therefore, we obtained three estimates of θ_i:

1. the MLE $\frac{y_i}{n_i}$;
2. the posterior mean of θ_i for the non-hierarchical Bayesian model;
3. the posterior mean of θ_i for the hierarchical Bayesian model.

				Posterior mean for	
i	y_i	n_i	MLE	non-hier.	hier.
1	0	15	0.000	0.059	*0.075*
2	18	148	0.122	0.127	*0.102*
3	1	10	0.100	**0.167**	**0.085**
	⋮			⋮	
10	8	97	0.082	0.091	*0.081*
11	29	256	0.113	0.116	*0.102*
12	24	360	0.067	0.069	*0.072*

NB: Compared with the non-hierarchical model, the hierarchical Bayesian model

- moved estimates towards the overall death rates, 0.077
- shrunk large estimates mostly for those hospitals with little data, ie small n_i

Hierarchical priors

We have specified a *hierarchical prior* for the surgical failure rates θ_i.

In general, suppose we have data $\mathbf{y}$ and parameters $\boldsymbol{\theta} = (\theta_1, ..., \theta_n)$

- Likelihood $p(\mathbf{y} \mid \boldsymbol{\theta})$ (1st level)
- Prior $p(\boldsymbol{\theta})$ depends on higher level parameter ϕ_2: $p(\boldsymbol{\theta} \mid \phi_2)$ (2nd level)

- $p(\phi_2)$ (3rd level)
  Marginal prior for $\boldsymbol{\theta}$ is then
  $$p(\boldsymbol{\theta}) = \int p(\boldsymbol{\theta} \mid \phi_2) p(\phi_2) d\phi_2$$

- We might add further levels
  $p(\phi_2 \mid \phi_3)$ (3rd level)
  ...
  $p(\phi_m)$ ($(m+1)$-th (top) level)
  Marginal prior for $\boldsymbol{\theta}$ is then
  $$p(\boldsymbol{\theta}) = \int \cdots \int p(\boldsymbol{\theta} \mid \phi_2) \times p(\phi_2 \mid \phi_3) \times \ldots \times p(\phi_{m-1} \mid \phi_m) \times p(\phi_m)\, d\phi_2 \ldots d\phi_m$$

- ϕ_k are called (kth level) *hyper-parameters*

- Theoretically there can be as many levels as necessary, but in practice it is usually hard to interpret parameters of level 3 or higher
- A non-informative prior is usually specified for the marginal distribution of the top-level parameters

For the hospital example:

$$
\begin{aligned}
Y_i \mid \theta_i &\sim \text{Binomial}(n_i, \theta_i) && \text{(Level 1)} \\
\theta_i \mid \alpha, \beta &\sim \text{Beta}(\alpha, \beta) && \text{(Level 2)} \\
\alpha &\sim \text{Exponential}(0.01) && \text{(Top level)} \\
\beta &\sim \text{Exponential}(0.01) && \text{(Top level)}
\end{aligned}
$$

3. Graphical models

Why graphical models?

Most realistic applications involves many interconnected random variables. We want an easy way

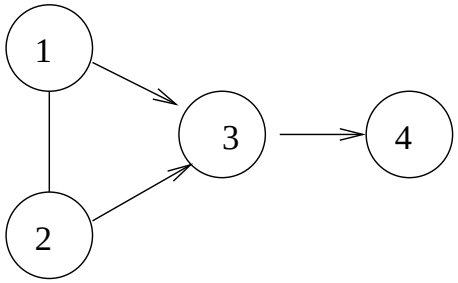
- to model and visualise the relationships between these random variables
- to figure out the properties of the model, e.g. conditional independence structure
- to simplify the fitting of the model

What is a graphical model?

A graphical model is

1. a *probability model* for multiple random variables
2. *(conditional) independence structure* of the model is characterised by a *graph*

Graphs



A graph consists of two sets: a set of nodes (or vertices) and a set of edges. Each edge connects a pair of nodes.

In a probabilistic graph: each node represents a random variable; the edges express association between these random variables.

Edges may be directed (arrows) or undirected (lines). A graph is called

- *undirected* if all its edges are undirected
- *directed* if all its edges are directed
- *mixed* if it contains directed and undirected edges

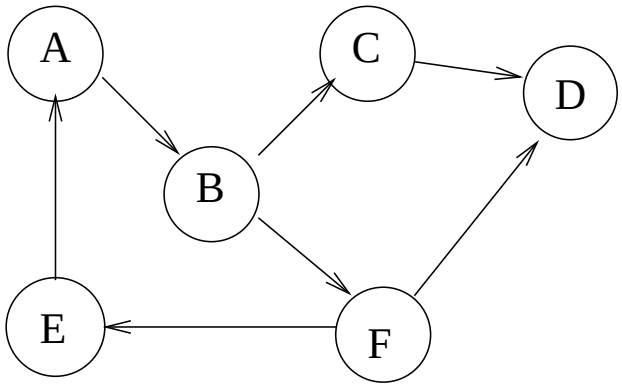
Undirected graphical models are also called *Markov random fields*.
Directed graphical models are also called *Bayesian networks*.

Directed Acyclic Graphs (DAGs)

What is a DAG?

A DAG is a directed graph that contains no directed cycles.

Is this a DAG?



Why DAGs?

DAGs are useful for visualising and investigating conditional dependence, e.g. causal relationship, between random variables.

How to build DAGs?

- Each random variable in the model is represented by a node
- An arrow pointing from one node to another indicates that the first variable 'causes' (influences) the second

Example

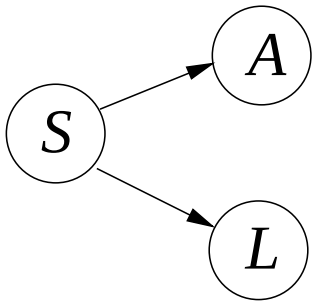
A = arterial disease (present/absent)

L = lung cancer (present/absent)

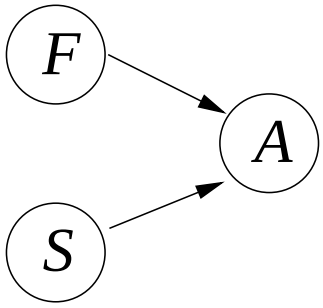
S = smoking amount (pack-years)

F = fat consumption (g per week)

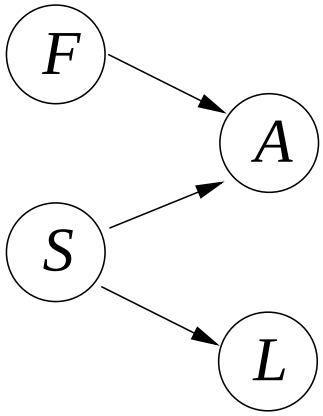
There is evidence smoking 'causes' (increases risk of) both lung cancer and arterial disease:



There is evidence that fat consumption is related to arterial disease:



Combining these two sub-models, we obtain:



Is

- F marginally independent of S?
- L and A are conditionally independent given S ($L \perp\!\!\!\perp A \mid S$)?

Example of conditional indep.

$X =$ height of child
$Y =$ mathematical ability of child
$Z =$ age of child

Is $X \perp\!\!\!\perp Y$?
Is $X \perp\!\!\!\perp Y \mid Z$?

The Factorisation Theorem

Write the random variables in a probability model as $X_1, \ldots, X_K$, say, and let $\mathbf{X} = (X_1, \ldots, X_K)$. If the model is represented by a DAG, the joint probability distribution, $p(\mathbf{X})$, of all the random variables, $\mathbf{X}$, in the model can be calculated using *Factorisation Theorem*:

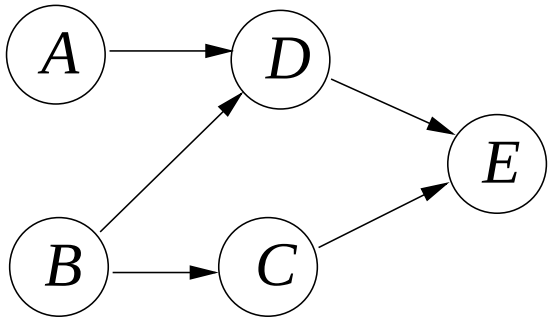
$$p(\mathbf{X}) = \prod_{k=1}^{K} p\left(X_k \mid \text{parents}\left[X_k\right]\right)$$

Note: We shall need this when we want to fit a Bayesian model.

Terminology for DAGs:

- *Parents* of a node are the nodes immediately 'upstream' (arrows point from parents)
- *Children* of a node are the nodes immediately 'downstream' (arrows point to children)

Example



What is the joint distribution of (A, B, C, D, E)?

$$\begin{aligned} p(A,B,C,D,E) &= p(A)p(B)p(C \mid B) \\ &\quad \times p(D \mid A,B)p(E \mid C,D) \end{aligned}$$

(since A and B have no parents; B is the only parent of $C \ldots$)

Full-conditional distributions

The *full-conditional distribution* of X_k is given by

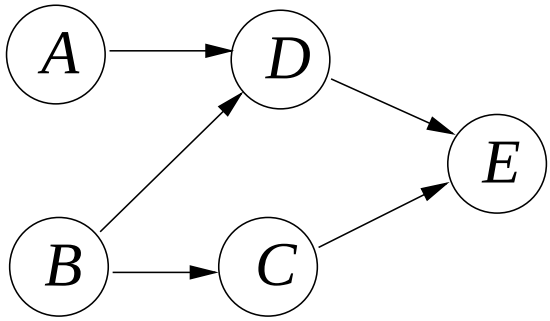
$$p(X_k \mid \mathbf{X}_{\setminus X_k}) \;\propto\; p(X_k \mid \text{parents}\,[X_k]) \times \prod_{W \,\in\, \text{children}[X_k]} p(W \mid \text{parents}\,[W]),$$

where $\mathbf{X}_{\setminus \mathbf{Y}}$ denotes the vector $\mathbf{X}$ excluding a sub-vector $\mathbf{Y}$.

Note:

- Full-conditional distributions are needed for fitting Bayesian models using MCMC methods

Example



What is the full-conditional distribution of C?

We have
$$\begin{aligned} \text{parents}\,[C] &= \{B\} \\ \text{children}\,[C] &= \{E\} \\ \text{parents}\,[E] &= \{C, D\} \end{aligned}$$

So, the full-conditional distribution of C is

$$p(C \mid A, B, D, E) \;\propto\; p(C \mid B) \times p(E \mid C, D) \;.$$

Note:

- The joint distribution of the model is

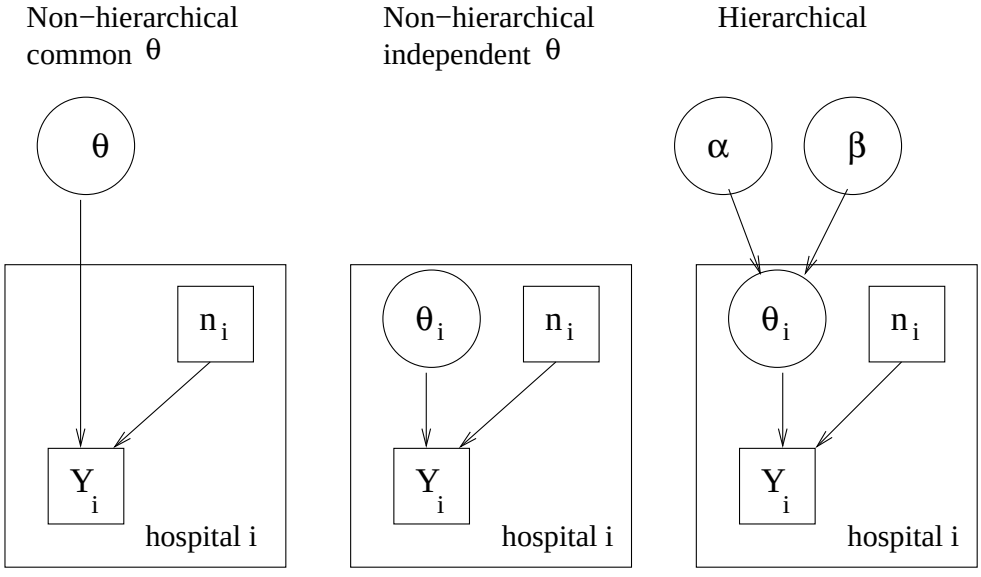
$$\begin{aligned} p(A, B, C, D, E) &= p(A)p(B)p(C \mid B) \\ &\quad \times p(D \mid A, B)p(E \mid C, D) \end{aligned}$$

4. Using DAGs for hierarchical models

DAGs can be used to represent hierarchical models. Conventionally, it uses

- circle nodes to represent unknown rvs (e.g. parameters, missing data)
- square nodes to represent known rvs (e.g. data)
- rectangular boxes to represent repetitive structures (e.g. one box for each hospital)

Our hospital models can be represented:



5. Summary

- *Hierarchical modelling involves breaking down the problem into layers and specifying a model for each layer*: a model for data given parameters; a model for parameters given hyper-parameters; maybe a model for hyper-parameters given higher-level hyper-parameters
- *It is useful when data obtained from similar-but-not-the-same units — parameters for different units are exchangeable*. Such models enables data on one unit to inform parameters of other units (borrowing strength). They move extreme estimates of units with little information towards population mean — this stabilises parameter estimates
- It is often difficult to specify informative priors for hyper-parameters, so we usually use non-informative (vague) priors for hyper-parameters

Obtaining marginal posterior distributions for parameters of a hierarchical model analytically is often not possible. We need MCMC.

Outline revisited

1. Non-hierarchical models
2. Hierarchical models (hierarchical priors)
3. Graphical models (DAG: Directed Acyclic Graph)
4. Using DAGs for hierarchical models
5. Summary

Next week: MCMC & WinBUGS

Please bring your laptop with WinBUGS installed