BCCS 2008/09: Graphical models and complex stochastic systems: Lecture 5: Hierarchical models

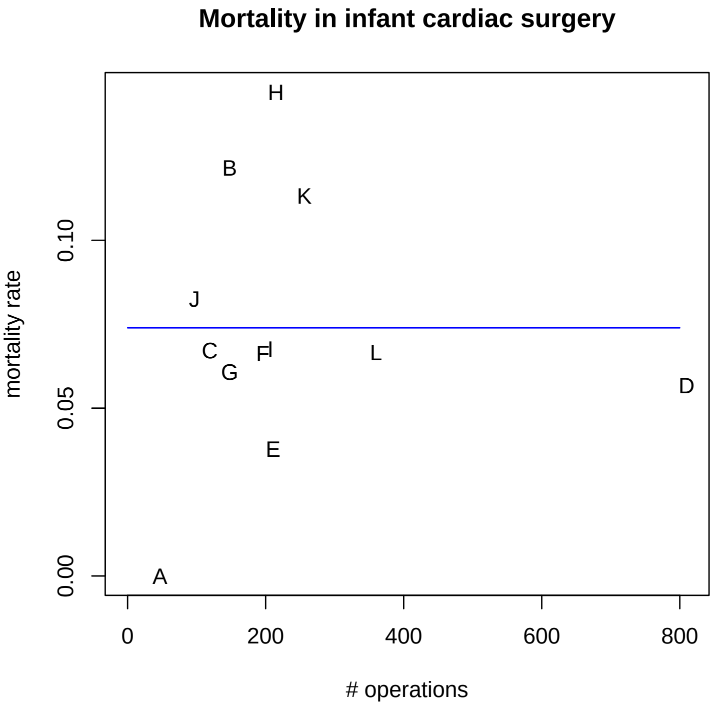
Now we are going to put DAGs to good use in statistical modelling.

5.1 Motivation for hierarchical modelling

How to make inference on multiple parameters $\{\theta_1, \ldots, \theta_I\}$ measured on I units (persons, centres, areas, ...) *which are related or connected by the structure of the problem ?*

The 'surgical' example

In 12 hospitals carrying out cardiac surgery on babies, the numbers of operations performed and mortality rates are recorded. What are the best and worst hospitals? Are the differences more than can be attributable to chance? What rate do you expect in the 13th hospital? Or in the 12th hospital, in a different year? In this example, θ_i is the true mortality rate in the ith hospital. Let Y_i and n_i be the number of deaths and the number of operations, in the ith hospital. We might assume $Y_i \sim \text{Binomial}(n_i, \theta_i)$.



We can identify three different assumptions:

1. **Identical parameters:** All the θ's are identical, in which case all the data can be pooled and the individual units ignored.

2. **Independent parameters:** All the θ's are entirely unrelated, in which case the results from each unit can be analysed independently (for example using a fully specified prior distribution within each unit)

   → individual estimates of θ_i are likely to be highly variable (unless very large sample sizes)

3. **Exchangeable parameters:** The θ's are assumed to be 'similar' in the sense that the 'labels' convey no information

Lessons from the 'surgical' data set

In the 12 hospitals, the 'raw' mortality rates vary between 0/47 (hospital A) and 31/215=0.1442 (H); the aggregated rate is 208/2814=0.0739. What are the 'true' rates in hospitals A and H?

Non-Bayesian answer 1. Assume that in hospital i, the number of deaths $Y_i \sim \text{Bin}(n_i, \theta)$. The maximum likelihood estimator of θ is $(\sum_i Y_i)/(\sum_i n_i) = 0.0739$, which applies to both A and H.

Non-Bayesian answer 2. Assume that in hospital i, the number of deaths $Y_i \sim \text{Bin}(n_i, \theta_i)$, independently. The maximum likelihood estimator of θ_i is $Y_i/n_i = 0$ for A and 0.1442 for H.

Could the θ_i all be equal? If θ is 0.0739, the chance that Y_{H} is as big or bigger than 31 is 0.000284. So, no!

Bayesian answer 1. Assume in addition that *a priori*, $\theta \sim \text{Beta}(\alpha, \beta)$ where α and β are say 4 and 46. (This gives a mean and variance for the Beta distribution roughly comparable to the sample mean and variance of the raw mortality rates). Then we get the posterior mean $= (\sum_i Y_i + \alpha)/(\sum_i n_i + \alpha + \beta) = 0.0740$ (for both A and H).

Bayesian answer 2. Making a similar prior assumption on each θ_i, the posterior mean of θ_i is $(Y_i + \alpha)/(n_i + \alpha + \beta) = 0.0412, 0.1321$ for A and H.

Which is best? Note that the Bayesian estimates are 'shrunk' towards the prior mean $\alpha/(\alpha + \beta) = 0.08$, to an extent depending on the 'denominator' n_i or n. This eliminates ridiculous conclusions like $\theta_{\text{A}} = 0$. However, it is still the case that only the data from hospital i is used in estimating θ_i. Surely the other hospitals' data carries information too? (For example, suppose that Y_{H} was missing: would you be able to guess its value better after having observed the other data?)

Our initial model 1 (Bayesian or non-Bayesian) revealed difficulty with the assumption that there was a common mortality rate θ in every hospital; we asked:

- Does this model adequately describe the random variation in outcomes for each hospital?
- Are the hospital failure rates more variable than our model assumes?

and concluded 'no' and 'yes', respectively.

Modelling the excess variation

Let's look at Bayesian model 2 above in more detail: we have modified model 1 to allow for a *different* failure probability, θ_i for each hospital i:

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

Interpretation:

- $\{\theta_i\}$, the 'true' surgical failure rate in the hospitals are viewed as a random sample from a common *population distribution*
  - ⇒ hospital failure rates are assumed to be **similar** but not identical
  - – Beta(α, β) prior describes the distribution of surgical failure rates amongst the 'population' of hospitals

How would you specify values for α and β?

Approximate 'empirical Bayes' approach

- Calculate crude failure rates y_i/n_i
- Calculate the observed mean and variance of the 12 values y_i/n_i
- Solve for $\widehat{\alpha}$ and $\widehat{\beta}$ to obtain a beta distribution with this mean and variance
- Using Beta($\widehat{\alpha}$, $\widehat{\beta}$) as a prior, apply Bayes theorem to obtain posteriors for true failure rates θ_i, $p(\theta_i|\widehat{\alpha}, \widehat{\beta}, y_1, y_2, \ldots, y_I)$

Potential problems with this approach:

- We are using the data twice:
  - Once to estimate the prior
  - Again to estimate θ_i for each hospital

  $\Rightarrow$ overestimate precision of our inference
- Using any point estimate for α and β ignores some posterior uncertainty about the population distribution of the θ_i's

5.2 Bayesian hierarchical models

The methods discussed here will allow us to do better, because we will be able to assume in advance that the true mortality rates across the hospitals are different (because the circumstances, patients, doctors, ... are different), but similar (because the operations, disease, ... are the same). The effect we will see is that the raw estimates are shrunk *towards each other.*

To do this, we need to deal with more than two sorts of variable – the parameters and data of ordinary Bayesian models. The hospitals problem has 3 levels of uncertainty – the hazard of this type of operation, the variability between hospitals, and chance factors in an individual patients' operation. Such models are called *hierarchical.*

- Assume a *joint probability model* for the entire set of parameters $(\theta_1, \theta_2, \ldots, \theta_I, \alpha, \beta)$ – requires us to assign known prior distributions to α, β, e.g.

$$\alpha \sim \text{Exponential}(1) \quad \text{and} \quad \beta \sim \text{Exponential}(1)$$

- Apply Bayes theorem to calculate the joint posterior distribution of all the unknown quantities simultaneously.

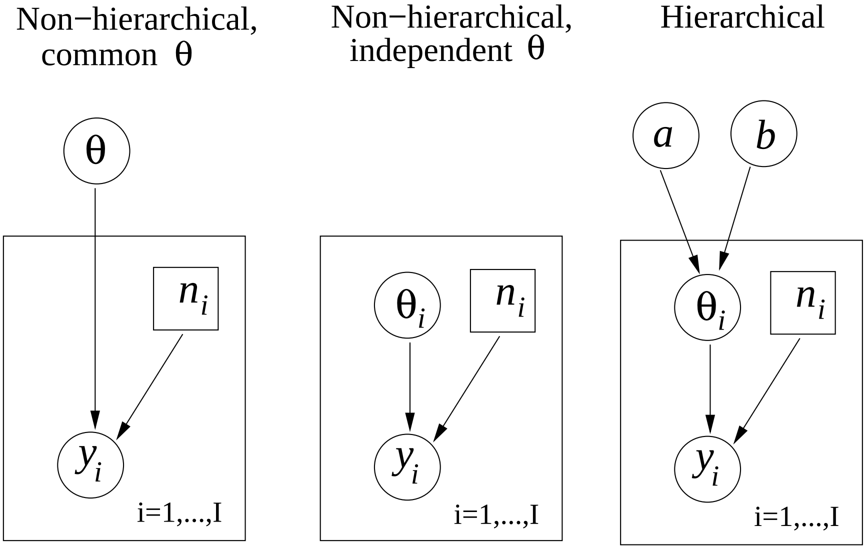
Level 1: $y_i \sim \text{Binomial}(n_i, \theta_i)$, independently for each i
Level 2: $\theta_i \sim \text{Beta}(\alpha, \beta)$, independently for each i
Level 3: Prior for α, β

Advantages of this approach

The posterior distribution for each θ_i

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

Such models are also called *Random effects* or *Multilevel* models.
Graphical models (DAGs) for surgical example:



5.3 Shrinkage and hierarchical models

To take a different example, suppose in each unit we observe a response x_i assumed to have a Normal likelihood

$$x_i \sim \mathrm{N}(\theta_i, \tau_i^2)$$

Unit means θ_i are assumed to be exchangeable, and to have a Normal distribution

$$\theta_i \sim \mathrm{N}(\mu, \sigma^2)$$

where μ and σ^2 are 'hyper-parameters', for the moment assumed known, as are τ_i^2.

It can be shown that, after observing x_i, Bayes' theorem gives

$$\theta_i | x_i \sim \mathrm{N}(w_i\mu + (1 - w_i)x_i, (1 - w_i)\tau_i^2)$$

where $w_i = \tau_i^2/(\tau_i^2 + \sigma^2) \in (0, 1)$ is the weight given to the prior mean.

A Bayesian model therefore leads to inferences for each θ_i giving intervals that are *narrower* than in the non-Bayesian approach, but *shrunk* towards the prior mean response. w_i controls both the 'shrinkage', and the reduction in the width of the interval: it depends on precision of the individual unit i relative to the variability between units. When $\{\tau_i^2\}$ are also given a prior, the same principles apply, although the solution is less explicit.

In a hierarchical model, μ and σ^2 are random, and the effect of this is more complicated again, and *best seen numerically*; the amount of shrinkage is not determined in advance – it is discovered from the data (an automatic consequence of Bayes' theorem). μ will also be shrunk towards the data in its posterior distribution, so that the θ_i are now shrunk towards a "typical" x value.

5.4 Exchangeability and de Finetti's theorem

'Exchangeability' is a formal expression of the idea that we find no systematic reason to distinguish the individual random variables $\theta_1, ..., \theta_I$ – a *judgement* that they are 'similar' but not identical.

An infinite sequence of 0/1 random variables $\theta_1, \theta_2, \ldots$ is called (infinitely) exchangeable if any finite subset has a joint distribution that is the same whatever the order in which the variables are written. E.g. $p(\theta_4, \theta_7, \theta_9) = p(\theta_7, \theta_9, \theta_4)$.

If the variables are independent Bernoulli(ϕ), they are obviously exchangeable. This remains true if ϕ is random (as in the coin-tossing example, with two biased coins), since e.g.

$$p(\theta_4, \theta_7, \theta_9) = \int_0^1 p(\phi)\phi^{\theta_4}(1-\phi)^{1-\theta_4}\phi^{\theta_7}(1-\phi)^{1-\theta_7}\phi^{\theta_9}(1-\phi)^{1-\theta_9}d\phi$$

(in the case ϕ has a continuous distribution), and this obviously only depends on $\theta_4+\theta_7+\theta_9$, not the order they appear. The remarkable thing is that the converse of this is true – the only way to get infinitely exchangeable 0/1 random variables is by Bernoulli trials with a fixed or random ϕ. This is (a form of) de Finetti's theorem. There are more general versions of the theorem, not just for 0/1 variables.

It gives mathematical support for using hierarchical models: if your prior beliefs about a set of parameters (e.g. the hospital mortality rates $\{\theta_i\}$) are exchangeable (really just a symmetry assumption), then without loss of generality you can model them as i.i.d. from some distribution given ϕ, and then make ϕ random.

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

Thus, under broad conditions an assumption of exchangeable units is mathematically equivalent to assuming the θ's are drawn at random from some population distribution.

5.5 What else do hierarchical models address?

Real data about real systems are complex: classic statistical methods are not enough. Among the features that real data might have that we could begin to handle are:

- repeated measures,
- heterogeneity between individuals,
- explanatory variables at individual and group level,
- measurement errors, multiple instruments,
- missing data, informative censoring,
- spatial or temporal structure.

5.6 Summary: why hierarchical?

Many interlinked arguments to favour the use of hierarchical models:

- by breaking down the problem in layers, able to separate structural judgments on observables, on parameters and subjective information

- reduces the arbitrariness of hyperparameter choice → “robustify” the inference
- natural structure for expressing dependence, prior correlations, ... in a plausible way (see next lectures)
- through shrinkage and borrowing of strength, parameter estimates are stabilised
- by de Finetti, if our beliefs are exchangeable, then they can be expressed mathematically by a hierarchical model.

5.7 Reading

Chapter 5 of Gelman et al, chapter 2 of Gilks et al.