Chapter 8

Hierarchical models

8.1 Introduction

Hierarchical models underpin a lot of real-world modelling. Sometimes we will have few or no data available that directly inform us about the things that we are interested in so borrow from available data that tell us about related things. This relates to the ideas in the previous chapter on DAGs because we will be interested in the flow of information from one set of observations to another.

A key text here is Bayesian Data Analysis by Gelman *et al.* (2013), and there are some useful observations in Kendall's Advanced Theory of Statistics Volume 2B: Bayesian Inference by O'Hagan and Forster (2004) (Chapters 6 and 7 in particular).

8.2 Exchangeability

Bruno De Finetti (1906–1985) was an early modern-statistician who wrote a two volume book that was grandly named "*Theory of Probability*". In that book, he laid out an axiomatic approach to probability that underpins modern day Bayesian statistics. A key idea formalised in his work was exchangeability.

A sequence of random variables is said to be *exchangeable* if we can reorder the sequence without changing the joint distribution of those random variables. Mathematically, we have a sequence $X_1, X_2, X_3, \dots$ and a bijective permutation operator $\sigma : \mathbb{N} \to \mathbb{N}$. If these random variables are exchangable, then

$$F_{X_1,X_2,X_3,\dots}(X_1, X_2, X_3, \dots) = F_{X_1,X_2,X_3,\dots}\left[X_{\sigma(1)}, X_{\sigma(2)}, X_{\sigma(3)}, \dots\right],$$

and *vice versa*. This may remind you of the modelling assumption that random variables are independent and identically distributed (i.i.d), and there is a strong relationship. In fact, independence implies exchangeability, the converse is not necessarily true.

Example 8.2.1

Imagine that I have selected six local councils and recorded each council's overspend in £m for the last financial year. Let us label these six unknowns as $x_1, \ldots, x_6$. What can you say about x_6, the overspend for the sixth council?

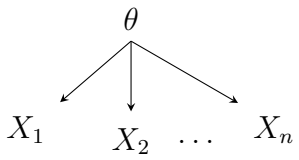
I have not given you any additional information so you should model them exchangeably, and each unknown value would be assigned an appropriate distribution over the real-line.

We now get the values for five of the six councils: 18, 5, 2, 7 and 8. A reasonable posterior predictive value for x_6 would be centred around 8 with a fair bit of spread, 0 to 25 perhaps.

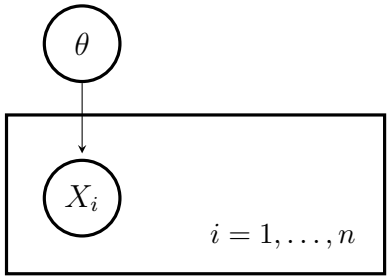
Changing the indices will not change this posterior estimate. Therefore, the x_i are exchangeable.

However, the x_i are certainly not independent: we would rightly assume that the overspend for the sixth council is similar to the observed spends.

Typically, we will talk about exchangeability of unknowns conditional on known parameter values (which will lead to the assumption of conditional independence).



An alternative representation that will be more useful later is:



Here, we have

$$X_i \perp X_j | \theta \quad \forall\, i \neq j.$$

8.3 Hierarchical models

Formally, a hierarchical model is a statistical model with multiple levels with each level representing a broader grouping of individual experimental or observed units. We will be using notation of the following form to capture the multiple levels for the units:

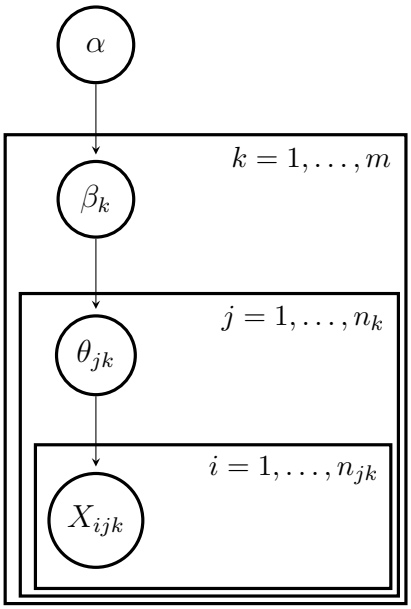
$$x_{ijk}$$

Here, the k indexes the overall group the unit belongs to, j indexes the subgroup and i indexes the observation within the subgroup.

These groups do not need to be symmetric in terms of the number of groups or units with those groups. This will lead us to having

$$n_{jk}$$

denoting sample size for the jth group within the ith overall group for a three-level hierarchy.

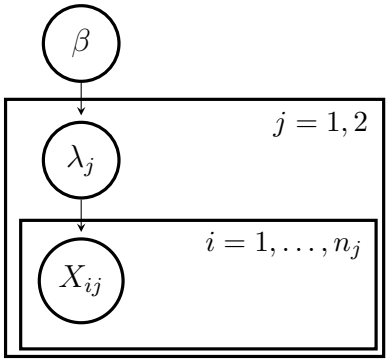


Note that we may model exchangeably at any of these level, and we could be modelling exchangeably at the highest level, but have strong dependencies at lower levels.

Example 8.3.1

Consider the following model:

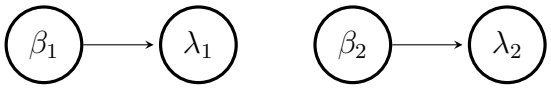
$$\begin{aligned} X_{ij}|\lambda_j &\sim \text{Poisson}(\lambda_j), \quad i = 1, \ldots, n_j, \\ \lambda_j|\beta &\sim \text{Exp}(\beta), \qquad j = 1, 2, \\ \beta &\sim \text{Exp}(1). \end{aligned}$$



We observe the following data:

$$\begin{aligned} n_1 = 3, \quad & \sum_{i=1}^{3} x_{i1} = 1, \\ n_2 = 10, \quad & \sum_{i=1}^{10} x_{i2} = 7. \end{aligned}$$

Now, suppose that we ignore the common β:



$$\begin{aligned} \lambda_1|x_{\bullet 1}, \beta_1 = 1 \sim \text{Gamma}(4, 2), \quad &\Longrightarrow \quad \text{Var}(\lambda_1|x_{\bullet 1}) \approx 0.127, \\ \lambda_2|x_{\bullet 2}, \beta_2 = 1 \sim \text{Gamma}(11, 8), \quad &\Longrightarrow \quad \text{Var}(\lambda_2|x_{\bullet 2}) \approx 0.067. \end{aligned}$$

Back to the full model, we have

$$\begin{aligned} \pi\left(\beta, \lambda_1, \lambda_2|x_{\bullet\bullet}\right) &\propto \pi(\beta)\pi(\lambda_1|\beta)\pi(\lambda_2|\beta)\pi(x_{\bullet 1}|\lambda_1)\pi(x_{\bullet 2}|\lambda_2) \\ &\propto \beta^2 \exp\left(-(\lambda_1 + \lambda_2 + 1)\beta\right) \lambda_1 \lambda_2^7 \exp\left(-3\lambda_1 - 10\lambda_2\right). \end{aligned}$$

Note that

$$\int_0^\infty \beta^2 \exp\left(-(\lambda_1 + \lambda_2 + 1)\beta\right) d\beta = \frac{\Gamma(3)}{(\lambda_1 + \lambda_2 + 1)^3}.$$

Therefore,

$$\pi\left(\lambda_1, \lambda_2|x_{\bullet\bullet}\right) \propto \frac{\lambda_1 \lambda_2^7 \exp\left(-3\lambda_1 - 10\lambda_2\right)}{(\lambda_1 + \lambda_2 + 1)^3}.$$

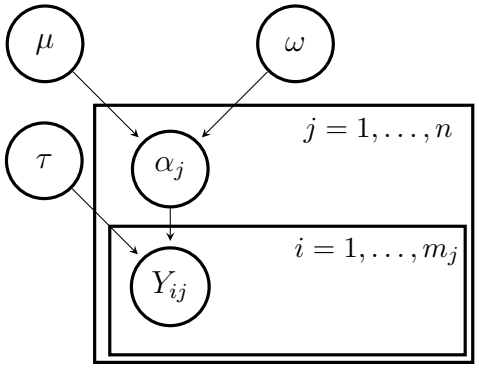
With the help of some numerical integration, we find that

$$\text{Var}(\lambda_1|x_{\bullet\bullet}) \approx 0.117 \text{ and } \text{Var}(\lambda_2|x_{\bullet\bullet}) \approx 0.065.$$

Example 8.3.2

Random-effects models are the archetypal hierarchical model:

$$\begin{aligned} Y_{ij}|\alpha_j, \tau &\sim \mathsf{N}(\alpha_i, \tau^{-1}), \quad i = 1, \ldots, m_j, \quad j = 1, .., n, \\ \alpha_j|\mu, \omega &\sim \mathsf{N}(\mu, \omega^{-1}), \quad j = 1, .., n, \\ \mu &\sim \mathsf{N}(0, 1), \\ \tau &\sim \mathsf{Gamma}(2, 1), \\ \omega &\sim \mathsf{Gamma}(1, 1). \end{aligned}$$



The influence of the prior distribution in general is to pull the likelihood towards the prior. When several parameters have a common prior mean, the posterior estimates will all be pulled towards the common mean. This means that Bayesian estimates of this nature will be less spread out than non-Bayesian estimates. This is known as *shrinkage*.

Example 8.3.3

For the random-effects model of **Example 8.3.2**, we can derive the following full conditional distribution for each α_j:

$$\alpha_j|y_{\bullet\bullet}, \alpha_{-j}, \mu, \tau, \omega \sim \mathsf{N}\left[\frac{\omega\mu + \tau\sum y_{ij}}{\omega + m_j\tau}, (\omega + m_j\tau)^{-1}\right].$$

A full conditional is not the same as the posterior, but it is a slice through it and we can see shrinkage in action for the group means α_j with the overall mean μ being used in an expression for the mean alongside the data-based estimate.

When tracking the influence of data on our posterior beliefs, it can be useful to quantify the relative reduction in variance. Define *resolution* to be:

$$R_x(\theta) = 1 - \frac{\mathsf{Var}(\theta|x)}{\mathsf{Var}(\theta)}.$$

If the resolution is 1, we have nothing left to learn about θ.

8.4 Probabilistic programming

Probabilistic programming is a form of programming that enables the (semi-)automation of Bayesian inference. It is different to normal programming in that it allows for the incorporation of uncertainty into the programming process. By using probabilistic programming, we can develop models that take into account the uncertainty of the data and make more informed decisions whilst taking advantage of the conditional independence structure in our models.

There are quite a few well established options that we could utilise:

- **BUGS**: Bayesian inference Using Gibbs Sampling — sampling from full conditionals,
- **JAGS**: Just Another Gibbs Sampler — cross-platform version of BUGS,
- **INLA**: Integrated Nested Laplace Approximation — approximate Bayesian inference for Markov random fields,
- **Stan**: named in honour of Stanislaw Ulam (1909-1984) — has several inbuilt algorithms for automated posterior sampling (plus tools for model selection).

The main algorithm that we will utilise in Stan is the no U-turns sampler (NUTS), which is a variant of Hamiltonian Monte Carlo (HMC). The HMC algorithm is an efficient MCMC algorithm that uses Hamiltonian dynamics to sample from a probability distribution. The simulation then follows the trajectories of the Hamiltonian dynamics allowing it to explore a large number of parameter values in a relatively short amount of time.

1. Initialise parameters
2. For each iteration,
    a) Choose a random momentum vector,
    b) Compute the Hamiltonian,
    c) Evolve the system using Hamiltonian dynamics,
    d) Compute the acceptance probability,
    e) Accept or reject.

HMC is particularly useful for exploring multi-modal distributions, which traditional MCMC methods struggle with. A really good explanation of the concepts involved in HMC and NUTS can be found here:

https://elevanth.org/blog/2017/11/28/build-a-better-markov-chain/

Stan model code is made up of three core components (this code is from **Example 8.3.1**):

```
data {
  int<lower=0> N;                  // num individuals
  int<lower=1> J;                  // num groups
  int<lower=1,upper=J> group[N];   // group for individual
  int<lower=0> X[N];               // observed random variables
}
parameters {
  real<lower=0> beta;          // hyperparameter
  real<lower=0> lambda[J];     // rate by group
}
model {
  // Prior
  beta ~ gamma(1, 1);
  for (j in 1:J)
      lambda[j] ~ gamma(1, beta);

  // Likelihood
  for (n in 1:N)
      X[n] ~ poisson(lambda[group[n]]);
}
```

Note that, in R, the model gets compiled into C++ so every time the model is changed R needs to compile again (which can take some time). This can be problematic if we are investigating changes to the prior parameters. A useful workaround is to include the prior parameter values in the data specification:

```
data {
  int<lower=0> N;                  // num individuals
  int<lower=1> J;                  // num groups
  int<lower=1,upper=J> group[N];   // group for individual
  int<lower=0> X[N];               // observed random variables
  real<lower=0> beta_shape;        // shape parameter for beta ~ Gamma(.,.)
  real<lower=0> beta_rate;         // rate parameter for beta ~ Gamma(.,.)
}
parameters {
  real<lower=0> beta;          // hyperparameter
  real<lower=0> lambda[J];     // rate by group
}
model {
  // Prior
  beta ~ gamma(beta_shape, beta_rate);
  for (j in 1:J)
```

```
      lambda[j] ~ gamma(1, beta);

  // Likelihood
  for (n in 1:N)
      X[n] ~ poisson(lambda[group[n]]);
}
```

There are four other useful elements that we can add to a Stan model to extend its utility:

- `transformed data`,
- `transformed parameters`,
- `generated quantities`,
- `functions`.

The R code for running a Stan model and producing posterior samples is relatively easy:

```
# load the library (note that you need Rtools installed)
library(rstan)

# compile the model
model <- stan_model(file = 'Hierarchical models/Poisson model.stan')

# generate a posterior distribution
fit <- sampling(model,
                chains = 4,
                iter = 100000,
                data = list(N = 3 + 10,
                            J = 2,
                            group = c(rep(1,3),
                                      rep(2,10)),
                            X = c(0,1,0,
                                  2,0,0,1,2,0,0,1,1,0)))
```

Then you can summarise the fits and extract the samples from the `stan fit` object:

```
# summary stats for all model parameters
summary(fit)

# extract the posterior samples
l1_samples <- extract(fit, pars = 'lambda[1]')$`lambda[1]`
```

```
l2_samples <- extract(fit, pars = 'lambda[2]')$`lambda[2]`

# plot histograms
hist(l1_samples)
hist(l2_samples)

# calculate variances
var(l1_samples)
var(l2_samples)
```

8.5 Priors

If a prior is not specified in Stan for a model parameter, θ say, then it will default to the following

$$\pi(\theta) \propto 1 \quad \forall\ \theta.$$

Example 8.5.1

We have

$$X|\mu \sim \text{Pareto}(\mu, 1),$$
$$\pi(\mu) \propto 1, \quad \mu > 0.$$

The posterior is then

$$\pi(\mu|X = x) \propto \frac{\mu}{x^2}.$$

We clearly have a problem because

$$\begin{aligned}\int_0^\infty \frac{\mu}{x^2} d\mu &= \lim_{t\to\infty} \int_0^t \frac{\mu}{x^2} d\mu \\ &= \lim_{t\to\infty} \frac{t^2}{2x^2},\end{aligned}$$

and the integral is undefined.

Apart from the challenge of potentially improper posterior distributions, we also have the issue of change of variables.

Example 8.5.2

Let's pretend that we know nothing about θ apart from it being positive:

$$\pi_\theta(\theta) \propto 1, \quad \theta > 0.$$

What do we know about $\phi = 1/\theta$?

$$\begin{aligned} \pi_\phi(\phi) &= F'_\theta\left(1/\phi\right) \\ &= \pi_\theta\left(1/\phi\right)\left|\frac{\partial(1/\phi)}{\partial\phi}\right| \\ &= \frac{1}{\phi^2}, \end{aligned}$$

which does not seem so flat.

8.6 Expert elicitation

We should incorporate existing knowledge into a Bayesian analysis through careful selection of a prior distribution.

8.6.1 Cromwell's rule

In a letter to the Church of Scotland, Cromwell stated "I beseech you, in the bowels of Christ, think it possible you may be mistaken". For us, Cromwell's rule amounts to never assigning probabilities of 0 or 1, which both signify certainty. If we are certain *a priori*, then no amount of evidence may change our minds.

Example 8.6.1

Consider a coin that is thought to be fair. One person says that they are sure the proportion of heads, θ, is 0.5 to 1 d.p ($\theta \sim$ Uni$(0.45, 0.55)$). Another person is even more certain, and they assign $\theta \sim Be(1000, 1000)$ ($\Pr(\theta > 0.55) = 4 \times 10^{-6}$).
Graph sketched in lecture.
The coin is tossed 500 times and 499 times we have heads.
1^{st} person: posterior $\Pr(\theta > 0.55|\text{data}) = 0$,
2^{nd} person: posterior $\Pr(\theta > 0.55|\text{data}) = 0.9999997$.

8.6.2 Elicitation methods

Throughout, we will assume we have some continuous, univariate parameter θ that we want to specify a prior distribution for.

People struggle to give judgements on statistical constructs like mean and variance. It is much more reliable to focus on:

- Mode (most likely value),
- Median (equal chance of being above or below),
- Percentiles (what value of θ is there 10% chance of being above?),
- Probabilities (what's the probability that $\theta > 0$?),
- Plausible ranges (give a range of plausible values, e.g. 6σ-rule).

8.6.3 The bisection method

Q1 Specify a value of θ such that there is an equal chance of the true value being above and below (θ_m).

Q2 Imagine this true value θ is definitely below θ_m. Specify a value of θ in range $(-\infty, \theta_m)$ such that there is an equal chance of the true value being above or below (θ_l).

Q3 Imagine this true value θ is definitely above θ_m. Specify a value of θ in range (θ_m, ∞) such that there is an equal chance of the true value being above or below (θ_u).

This method is all about judging equal chances, which is easier than specifying probabilities or quartiles directly.

Example 8.6.2

What is my sister's height?

Having the three quartiles does not specify a distribution uniquely, and we may need to make a judgement about a suitable probability distribution.

If we are fitting a normal distribution to judgements (θ_l θ_m θ_u), we might use:

- $\theta_m = \mu$,

- $\theta_l = \mu - 0.6745\sigma$,
- $\theta_u = \mu + 0.6745\sigma$.

If we cannot determine μ and σ^2 exactly, we can choose μ and σ^2 that minimises:

$$(\theta_l^* - \theta_l)^2 + (\theta_m^* - \theta_m)^2 + (\theta_u^* - \theta_u)^2.$$

Here θ_l is the true lower quartile from $N(\mu, \sigma^2)$.

Example 8.6.3

My sister's height continued.

- $\theta_m^* = 5'9" = 175cm$
- $\theta_l^* = 5'7" = 170cm$
- $\theta_u^* = 5'11" = 180cm$

Use website:
https://jeremy-oakley.shinyapps.io/SHELF-single/

Here are the key principles of elicitation:

I Transparency.

II Ask questions that the expert can answer - training, flexibility.

III Question the expert $\rightarrow$ Fit distribution $\rightarrow$ Feedback distribution $\rightarrow$ are revisions needed? (if yes: back to start)(if no: use this in analysis).

IV Whatever prior is chosen, we should consider sensitivity of our analysis to that choice.

8.6.4 What is my prior worth?

To consider the influence of the prior we could ask the following: what number of observations is our prior worth?

Example 8.6.4

$x|\theta \sim \text{Bin}(n, \theta)$ and $\theta \sim \text{Be}(\alpha, \beta)$: $\theta|X = x \sim \text{Be}(\alpha + x, \beta + n - x)$. By properties of the beta distribution, $\text{E}(\theta) = \frac{\alpha}{\alpha+\beta}$ and $\text{E}(\theta|x) = \frac{\alpha+x}{\alpha+\beta+n}$.
We can rewrite:

$$\text{E}(\theta|x) = \frac{(\alpha + \beta)\text{E}(\theta)}{\alpha + \beta + n} + \frac{n\hat{\theta}}{\alpha + \beta + n}$$

where $\hat{\theta} = \frac{x}{n}$.
This is a weighted average of $\text{E}(\theta)$ and the data estimate, $\hat{\theta}$. In this average, the prior is worth $(\alpha + \beta)$ and the data estimate is worth n.