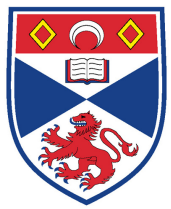
MT4531/5731: (Advanced) Bayesian Inference

Hierarchical models

Nicolò Margaritella

School of Mathematics and Statistics, University of St Andrews

–



Outline

1 Introduction

2 Example - California earthquakes

Outline

1 Introduction

2 Example - California earthquakes

Hierarchical models (1)

- The essential idea of hierarchical modelling is to 'learn' the prior to use for the data we are analysing, by looking at related data sets.
- These models can be fitted easily within a Bayesian framework via data augmentation.

Hierarchical models (2)

Hierarchical models extend what we've done so far:

- There are j groups for our observations y_{ji}.
- The behaviour of all groups is supposed to be similar (but not equal).
- The parameters θ_j for every group come from a common probability distribution (whose parameters need to be estimated) that can be considered a **Bayesian random effect**.

Hierarchical models (3)

Examples:

- In the rats example discussed in previous lectures, we defined rat_{ji}= weight of rat j at time i. We want to account for dependence between observations of a certain rat j.
- $y_{j,i}$= household income in city j and household i in Scotland. Household incomes within a city vary less that between cities.
- $y_{j,i}$= indicator that infant has low birth weight ($<$2000mg) for infant i born in hospital j.
  The probability that a newly born infant has a low birth weight differs between hospitals.
- etc...

Hierarchical models (4)

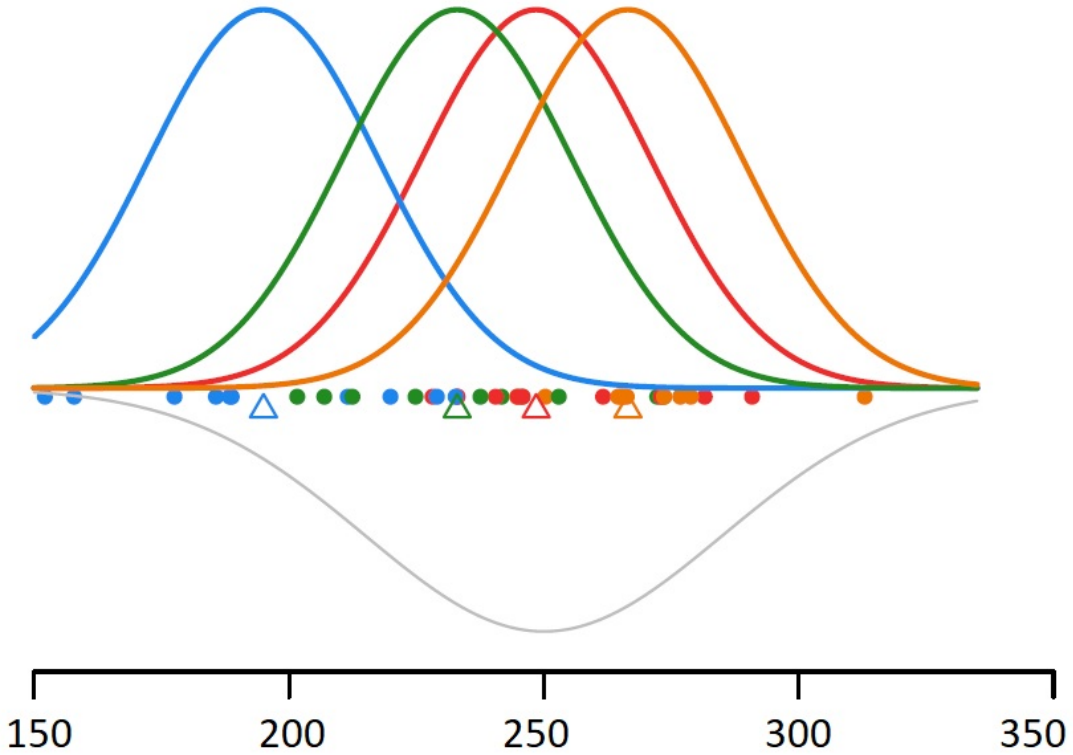
- Recall the rats model: Let Y_{ji} denote the weight of rat $j = 1, \ldots, 30$ at time $i = 1, \ldots, 5$. We assume a very simple linear model, where weight is linearly regressed on time,

$$Y_{ji} \sim N(\theta + \beta z_i, \sigma^2),$$

- α, β and σ^2 are parameters to be estimated; and z_i denotes the ith (normalised) time.

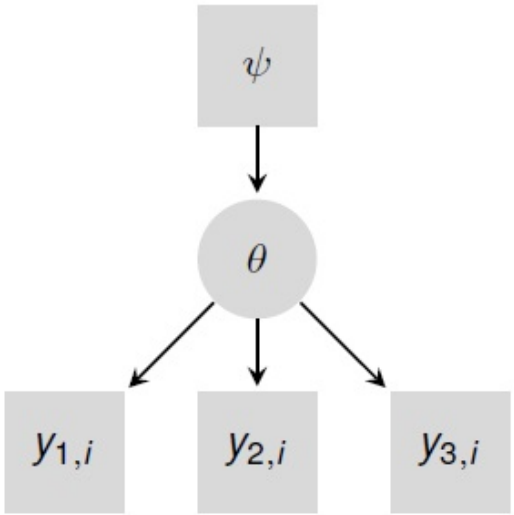
Hierarchical models (5)

- It might be sensible to assume that each rat's weight has its own intercept θ_j (coloured triangles). Furthermore, these θ_js can be considered as draws from a common probability distribution with mean ψ (grey density)



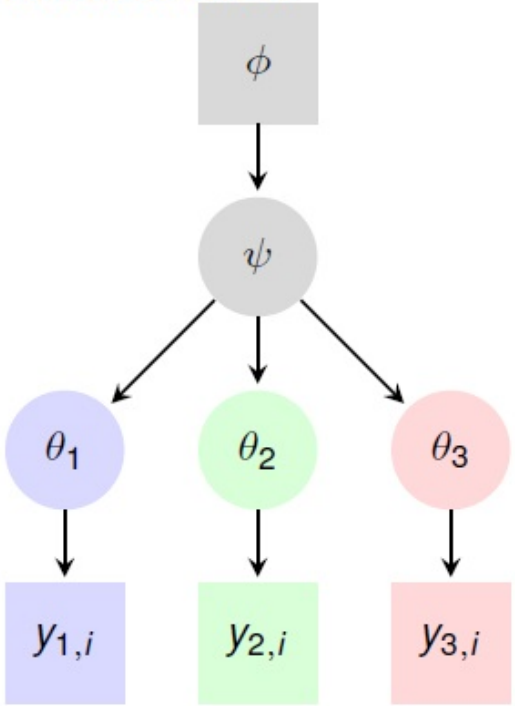
Hierarchical models (6)

Common effect



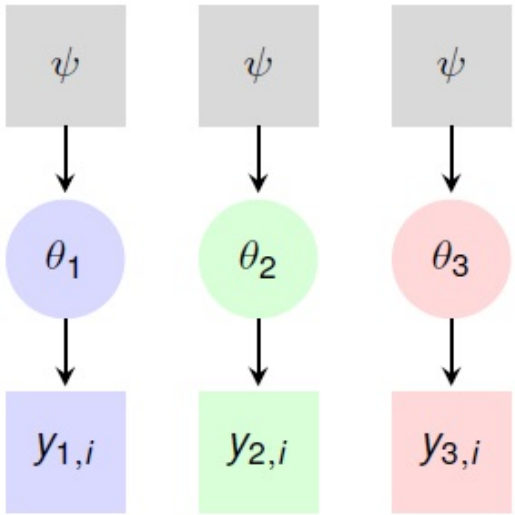
$$y_{j,i}|\theta \sim dist(\theta)$$
$$\theta|\psi \sim dist(\psi)$$

Random effect



$$y_{j,i}|\theta_j \sim dist(\theta_j)$$
$$\theta_j|\psi \sim dist(\psi)$$
$$\psi|\phi \sim dist(\phi)$$

Fixed effect



$$y_{j,i}|\theta_j \sim dist(\theta_j)$$
$$\theta_j|\psi \sim dist(\psi)$$

Hierarchical models (7)

- A random effect model (hierarchical model) for the rats example is

$$Y_{ji} \sim N(\theta_j + \beta z_i, \quad \sigma^2)$$
$$\theta_j \sim N(\psi, \kappa)$$

- Important: in order to be a hierarchical model, ψ and κ should also be random quantities that we estimate with the data; i.e. $\psi, \kappa \sim p(\psi, \kappa|\phi)$.

Hierarchical models (8)

- Note that there is a hierarchy to this model, hence the name.
- The hierarchy can be viewed from the observed outcomes up:

$$\begin{array}{ccccc} \text{Data \& Likelihood} & & \text{Prior} & & \text{Hyperprior} \\ f(y|\theta,\beta) & \Rightarrow & g(\theta|\psi,\kappa) & \Rightarrow & h(\psi,\kappa|\phi) \end{array}$$

- More levels can be imagined: person $\rightarrow$ household $\rightarrow$ city $\rightarrow$ country
- The model for the observations can include both hierarchical and common effects; e.g.,

$$y_{j,i} \sim \text{Normal}\left(\theta_j + \beta z_i, \sigma^2\right)$$

where $\theta_j \sim \text{Normal}\ (\psi, \kappa)$ and β is, e.g., $N(0, 10^5)$.

- This is sometimes called a *Mixed Effects* model.

Hierarchical models (9)

- In our rat example, the random effect θ_j is used solely to generate dependence between the is observations of a rat j. Hence θ_j are considered auxiliary random quantities which are added to the model to generate dependence within groups.
- Thus, the θ_j parameters are of no interest themselves. This implies that the model parameters are, in fact, ψ, β, σ^2 and κ. The likelihood of this random effects model can be given by,

$$f(\boldsymbol{y}|\psi,\beta,\sigma^2,\kappa) = \int f(\boldsymbol{y},\boldsymbol{\theta}|\psi,\beta,\sigma^2,\kappa)d\boldsymbol{\theta}$$
$$\int f(\boldsymbol{y}|\psi,\beta,\sigma^2,\kappa,\boldsymbol{\theta})f(\boldsymbol{\theta}|\psi,\beta,\sigma^2,\kappa)d\boldsymbol{\theta}.$$

- Note that the $\boldsymbol{\theta} = \{\theta_j : j = 1,\ldots,30\}$ could be integrated out within the likelihood. However, we can adopt a data augmentation approach, keeping the θ_j in our model to simplify likelihood calculations.

Hierarchical models (10)

- In this example, we have that,

$$f(\boldsymbol{y}|\psi,\beta,\sigma^2,\kappa,\boldsymbol{\theta}) = f(\boldsymbol{y}|\beta,\sigma^2,\boldsymbol{\theta})$$
$$= \prod_{j=1}^{N}\prod_{i=1}^{T} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_{ji}-\theta_j-\beta z_i)^2}{2\sigma^2}\right),$$

since $\boldsymbol{y}$ is independent of ψ and κ given $\boldsymbol{\theta}$. Note that,

$$f(\boldsymbol{\theta}|\psi,\beta,\sigma^2,\kappa) = f(\boldsymbol{\theta}|\psi,\kappa)$$
$$= \prod_{j=1}^{N} \frac{1}{\sqrt{2\pi\kappa}} \exp\left(-\frac{(\theta_j-\psi)^2}{2\kappa}\right),$$

since $\boldsymbol{\theta}$ is independent of σ^2 and β, given ψ and κ.

Hierarchical models (11)

- In general, for random effect models the necessary integration is analytically intractable (in this case the integration is analytically tractable).
- Within the Bayesian framework a data augmentation can be implemented: we treat the $\boldsymbol{\theta}$ as "auxiliary variables" and form the joint posterior distribution of the model parameters $\boldsymbol{\alpha} = (\psi, \beta, \sigma^2, \kappa)$ and auxiliary variables, $\boldsymbol{\theta}$,

$$\pi(\boldsymbol{\alpha}, \boldsymbol{\theta}|\boldsymbol{y}) \propto f(\boldsymbol{y}|\beta, \sigma^2, \boldsymbol{\theta}) f(\boldsymbol{\theta}|\psi, \kappa) p(\boldsymbol{\alpha}).$$

To obtain inference on the model parameters $\boldsymbol{\alpha}$, sample from this joint posterior distribution (using MCMC) and marginalise to obtain estimates of the posterior summary statistics of interest.

Outline

1 Introduction

2 Example - California earthquakes

California earthquakes (1)

- Now we look at an example where the random effects take a central role in the analysis.
- Suppose we are interested in predicting the occurrence of large earthquakes on a particular fault in southern California. On this particular fault, only 2 large earthquakes have been recorded in the last 12 years. The inter-event times recorded are:

$$Y_1 = \{4.75, 5.52\}$$

- We assume the inter-events are Exponential(λ_1). To predict future earthquakes, we need to estimate λ_1.

California earthquakes (2)

- There are also many other faults in Southern California that also have earthquakes. Suppose we have data from K other faults in total, with Y_k denoting the inter-arrival times on fault K.
- **Hierarchical model idea:** We want to use the data from faults $Y_2, Y_3, \ldots, Y_K$ to help us better estimate the parameter that generated the event times λ_1 on the original fault which we are interested in.
- This is not subjective since all prior information is coming from other data sets. But it is not objective either since we are analysing the current data set using outside information.

California earthquakes (3)

- Define Y_{ji} the i^{th} inter-arrival time on fault j and assume $Y_{ji} \sim \text{Exp}(\lambda_j)$. Let's look at different possible analyses:
- **No pooling** - we assume that there is no information that we can share across faults.
- **Complete pooling** - An alternative approach is to assume that every fault is exactly the same, so that all inter-event times have the same distribution i.e. $\lambda_1 = \lambda_2 = \cdots = \lambda_K$.

$\Rightarrow$ Neither approach is ideal. Hierarchical modelling aims at mediating between the two extremes.

California earthquakes (4)

- $\Rightarrow$ **Hierarchical model basic idea**: we choose a prior for $\lambda_{j=j'}$ using the information from the other datasets.
- $\Rightarrow$ If we have no obvious reason to think our fault is different from the other faults, we can combine the no-pooling and complete-pooling approaches by using the other faults to set the parameters of the prior distribution for our fault.
- $\Rightarrow$ Unlike no pooling, we are taking the other faults into account. But unlike complete pooling, we are not assuming our fault is identical (i.e. same parameter value) to any other faults.

California earthquakes (5)

- A hierarchical model for our example is

$$Y_{ji} \sim \text{Exp}(\lambda_j)$$
$$\lambda_j \sim \Gamma(a, b)$$
$$a \sim \text{U}(0, \infty); \;\; b \sim \text{U}(0, \infty)$$

$\Rightarrow$ If we have no obvious reason to think our fault is different from the other faults, we can assume that each of the values of λ_j are independent samples from a known distribution (a Gamma distribution in our case).

$\Rightarrow$ Key point: we treat a and b as extra parameters to be estimated along with the λ_js. Our unknown parameter vector is: $\boldsymbol{\theta} = (\lambda_1, \ldots, \lambda_K, a, b)$.

- This model might seem complex but we can easily sample from the posterior of all parameters.

$\Rightarrow$ see further demonstration in class.

Try at home

- In Exercise 3 in Tutorial 7 you are challenged to code the California's earthquakes example in NIMBLE and run the complete analysis. By completing it, you will have run your first advanced Bayesian model!