

Chapter 6

Hierarchical models

Hierarchical models

- ▶ Hierarchical modeling provides a framework for building complex and high-dimensional models from simple and low-dimensional building blocks
- ▶ Of course, it is possible to analyze these models using non-Bayesian methods
- ▶ However, this modeling framework is popular in the Bayesian literature because MCMC is conducive to hierarchical models
- ▶ Both “divide and conquer” big problems by splitting them into a series of smaller problems in the same way

We build models!

1D Statistician's creation

	M	A	P	S
5	P	O	L	L
6	A	D	I	E
7	D	E	V	A
8	S	L	E	

ACROSS

- 1 Google service with a "street view"
- 5 Spanish for "chicken"
- 6 Something to bid while leaving
- 7 ___ Patrick, 2020 presidential candidate
- 8 A saucer is a round one

DOWN

- 1 Statistician's creation
- 2 People's Sexiest Man ___
- 3 Beg
- 4 Genre for Otis Redding and Tina Turner
- 5 Wear for a football player

Outline

These notes cover Chapter 6

- ▶ Building a hierarchical model through layers
- ▶ Directed acyclic graphs
- ▶ Several examples

Hierarchical models

Often Bayesian models can be written in the following layers of the hierarchy

1. **Data layer:** $[Y|\theta, \alpha]$ is the likelihood for the observed data Y given the model parameters
2. **Process layer:** $[\theta|\alpha]$ is the model for the parameters θ that define the latent data generating process
3. **Prior layer:** $[\alpha]$ prior for hyperparameters

Epidemiology example - Data layer

- ▶ Let S_t and I_t be the number of susceptible and infected individuals in a population, respectively, at time t
- ▶ The data Y_t is the number of observed cases at time t
- ▶ The data layer models our ability to measure the process I_t
- ▶ **Data layer:** $Y_t|I_t \sim \text{Binomial}(I_t, \rho)$
- ▶ This assumes no false positives and false negative probability ρ

Epidemiology example - Process layer

- ▶ Scientific understanding of the disease is used to model disease propagation
- ▶ We might select the simple Reed-Frost model

Process layer:

$$I_{t+1} \sim \text{Binomial} \left[S_t, 1 - (1 - q)^{I_t} \right]$$
$$S_{t+1} = S_t - I_{t+1}$$

- ▶ This assumes all infected individuals are removed from the population before the next time step
- ▶ Also that q is the probability of a non-infected person coming into contact with and contracting the disease from an infected individual

Epidemiology example - Prior layer

- ▶ The epidemiological process-layer model expresses the disease dynamics up to a few unknown parameters
- ▶ The Bayesian model is completed using priors, say,
- ▶ **Prior layer:**

$$\begin{aligned}I_1 &\sim \text{Poisson}(\lambda_1) \\S_1 &\sim \text{Poisson}(\lambda_2) \\p, q &\sim \text{beta}(a, b)\end{aligned}$$

When to stop adding layers?

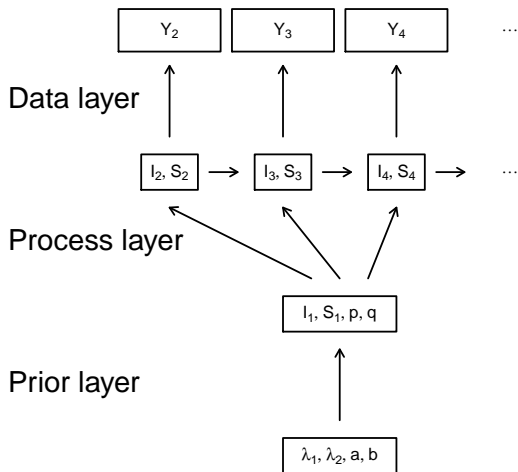
- ▶ In the previous example a , b , λ_1 and λ_2 are fixed
- ▶ But we will have uncertainty about the correct value
- ▶ Maybe replace a fixed value with another layer, say $a \sim \text{Uniform}(0, \theta)$?
- ▶ Then maybe $\theta \sim \text{Exponential}(\xi)$, $\xi \sim \text{Uniform}(0, \eta)$, etc.
- ▶ Rule of thumb: Be careful assigning priors to parameters in layers without replication.
- ▶ For example, even if we knew p exactly this would be just one value and we couldn't hope to estimate the parameters of its beta distribution.

Directed acyclic graphs (DAGs)

- ▶ A DAG is a graphical representation of a hierarchical model
- ▶ DAGS sometimes go by the name Bayesian networks
- ▶ Each observation and parameter is a node
- ▶ An arrow for X to Y means that the conditional distribution of Y depends on X
- ▶ “Directed” means that arrows only go one way
- ▶ Acyclic means there are no cycles, e.g.,

$$X \rightarrow Y \rightarrow Z \rightarrow X$$

Epidemiology example - DAG



Directed acyclic graphs (DAGs)

- ▶ Building models this way ensures we will always have a valid joint distribution
- ▶ For example, say we need to specify the joint distribution of (X, Y, Z)
- ▶ Any joint distribution can be written as

$$f(X, Y, Z) = f(X)f(Y|X)f(Z|X, Y)$$

- ▶ This is a fully-connected DAG
- ▶ Ad-hoc constructions like

$$f(X, Y, Z) = f(X|Z)f(Y|X)f(Z|X, Y)$$

may or may not give a valid joint PDF

Hierarchical models and MCMC

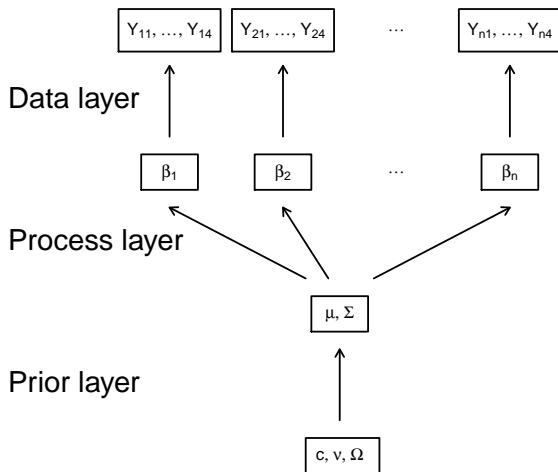
- ▶ Consider the classic one-way random effects model:

$$Y_{ij} \sim N(\theta_i, \sigma^2) \quad \text{and} \quad \theta_i \sim N(\mu, \tau^2)$$

where Y_{ij} is the j^{th} replicate for unit i and $\alpha = (\mu, \sigma^2, \tau^2)$ has an uninformative prior

- ▶ This hierarchy can be written using a directed acyclic graph

Random effects example - DAG



Hierarchical models and MCMC

- ▶ MCMC is efficient in this case even if the number of parameter or levels of the hierarchy is large
- ▶ You only need to consider “connected nodes” when you update each parameter
- ▶
 1. $[\theta_i | \cdot]$
 2. $[\mu | \cdot]$
 3. $[\sigma^2 | \cdot]$
 4. $[\tau^2 | \cdot]$
- ▶ Each of these updates is a draw from a standard one-dimensional normal or inverse gamma

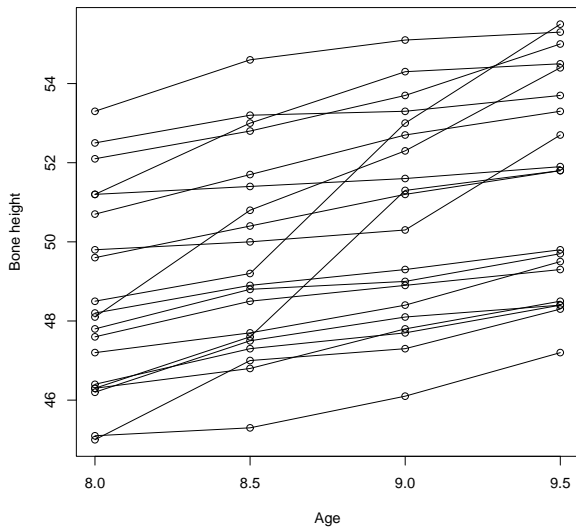
Random slopes model

- ▶ Let Y_{ij} be the j^{th} observation for subject i
- ▶ As an example, consider the data plotted on the next slide were Y_{ij} is the bone density for child i at age X_j .
- ▶ Here we might specify a different regression for each child to capture variability over the population of children:

$$Y_{ij} \sim \text{Normal}(\gamma_{0i} + X_j \gamma_{1i}, \sigma^2)$$

- ▶ $\gamma_i = (\gamma_{i0}, \gamma_{i1})^T$ controls the growth curve for child i
- ▶ These separate regression are tied together in the prior, $\gamma_i \sim \text{Normal}(\beta, \Sigma)$, which borrows strength across children
- ▶ This is a linear mixed model: γ_i are random effects specific to one child and β are fixed effects common to all children

Bone height data



Draw a DAG and work out the full conditional for γ_1

Missing data models

- ▶ We will deal with missing data in the linear regression context, but the ideas apply to all models

- ▶ The model is

$$Y_i \sim \text{Normal}(\beta_0 + \beta_1 X_{i1} + \dots + \beta_p X_{ip}, \sigma^2)$$

- ▶ Often either Y_i or elements X_{ij} are missing
- ▶ We will study separately the case of missing responses and missing covariates

Missing responses

- ▶ If the response is missing this is essentially a prediction problem
- ▶ We have seen how to handle this in JAGS
- ▶ We obtain samples from the PPD of Y_i
- ▶ At each MCMC iteration we simply draw

$$Y_i \sim \text{Normal}(\beta_0 + \beta_1 X_{i1} + \dots + \beta_p X_{ip}, \sigma^2)$$

- ▶ This distribution accounts for random error as well as uncertainty in the model parameters
- ▶ For the other updates the data are essentially complete
- ▶ If only responses are missing, can we delete them for the purpose of estimating β ?

Missing covariates

- ▶ Now say all responses are observed, but a some covariates are missing
- ▶ The simplest approach is imputation, e.g., just plug in the sample mean of the covariate for the missing values
- ▶ This doesn't account for uncertainty in the imputations
- ▶ Bayesian methods handle this well using MCMC

Missing covariates

- ▶ The main idea is to treat the missing values as unknown parameters in the Bayesian model
- ▶ Unknown parameters need priors, so missing $\mathbf{X}_i = (X_{i1}, \dots, X_{ip})^T$ must have priors such as

$$\mathbf{X}_i \sim \text{Normal}(\mu_X, \Sigma_X)$$

- ▶ Assumptions about missing data:
 - ▶ Missing status is independent of Y and \mathbf{X}
 - ▶ Covariates are Gaussian
- ▶ There are ways to relax both assumptions, but it becomes complicated

Missing covariates

- ▶ Of course if the prior is way off, the results will be invalid
- ▶ For example, if in reality the data are not missing at random the Bayesian model will likely give bad results
- ▶ Example of non-random missingness:
 - ▶ If specified correctly, the model will lead to inference for β that properly accounts for uncertainty about the missing data

Hierarchical linear regression model with missing data

- ▶ $Y_i | \mathbf{X}_i, \beta, \sigma^2 \sim \text{Normal}(\mathbf{X}_i^T \beta, \sigma^2)$
- ▶ $\mathbf{X}_i | \boldsymbol{\mu}, \Sigma \sim \text{Normal}(\boldsymbol{\mu}, \Sigma)$
- ▶ $p(\beta) \propto 1$
- ▶ $\sigma^2 \sim \text{InvG}(0.01, 0.01)$
- ▶ $\boldsymbol{\mu} \sim \text{Normal}(0, 100^2 I_p)$
- ▶ $\Sigma \sim \text{InvWishart}(0.01, 0.01 I_p)$

If some observations have missing Y and some have missing X , can we delete those with missing Y ? Can we delete those with missing X ?

Overview of the Gibbs sampling algorithm

- ▶ The full conditional of missing Y_i is:

- ▶ The full conditional of missing X_i is:

- ▶ In fact, all the full conditionals are conjugate

Worked examples

The course website includes three complete data analyses of hierarchical models

- ▶ Missing data analysis of 2016 Boston marathon data
- ▶ Analysis of tyrannosaurid growth curves
- ▶ Species distribution mapping via data fusion