(8) Hierarchical models

ST440/540: Applied Bayesian Statistics
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.
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 \(Y\)

2. **Process layer**: \([\theta|\alpha]\) is the model for the parameters \(\theta\) 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**: $Y_t | I_t \sim \text{Binomial}(I_t, p)$

- This assumes no false positives and false negative probability $p$
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
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:**

\[ I_1 \sim \text{Poisson}(\lambda_1) \]
\[ S_1 \sim \text{Poisson}(\lambda_2) \]
\[ p, q \sim \text{beta}(a, b) \]
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 (DAG; also called Bayesian network or belief network)

Epidemiology example - DAG

Data layer

Process layer

Prior layer

ST440/540: Applied Bayesian Statistics (8) Hierarchical models
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
Two-way random effects model

- [http://www4.stat.ncsu.edu/~reich/ABA/code/Ozone](http://www4.stat.ncsu.edu/~reich/ABA/code/Ozone)

- The ozone measurement at spatial location \(i\) and day \(j\) is denoted \(Y_{ij}\)

- We fit the model

\[
Y_{ij} \sim \text{Normal}(\mu + \alpha_i + \gamma_j, \sigma^2)
\]

- \(\mu\) is the overall mean

- \(\alpha_i\) is the random effect for location \(i\)

- \(\gamma_j\) is the random effect of day \(j\)
Two-way random effects model

- Model: \( Y_{ij} \sim \text{Normal}(\mu + \alpha_i + \gamma_j, \sigma^2) \)

- Priors for the fixed effects model:
  \[ \alpha_i \sim \text{Normal}(0, 10^2) \quad \text{and} \quad \gamma_j \sim \text{Normal}(0, 10^2) \]

- Priors for the random effects model:
  \[ \alpha_i \sim \text{Normal}(0, \sigma^2_a) \quad \text{and} \quad \gamma_j \sim \text{Normal}(0, \sigma^2_g) \]

- These look similar, when do we say something is a random effect in a Bayesian analysis?
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:

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

- $\gamma_i = (\gamma_{i0}, \gamma_{i1})^T$ 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: $\gamma_i$ are random effects specific to one child and $\beta$ are fixed effects common to all children.
Bone height data

---

ST440/540: Applied Bayesian Statistics (8) Hierarchical models
Prior for a covariance matrix

- The random-effects covariance matrix is \( \Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix} \)

- \( \sigma_1^2 \) is the variance of the intercepts across children

- \( \sigma_2^2 \) is the variance of the slopes across children

- \( \sigma_{12} \) is the covariance between the intercepts and slopes

- Prior 1: \( \sigma_1^2, \sigma_2^2 \sim \text{InvGamma} \) and \( \rho = \frac{\sigma_{12}}{\sigma_1 \sigma_2} \sim \text{Unif}(−1, 1) \)

- Prior 2: Inverse Wishart works better in higher dimensions
The inverse Wishart distribution is the most common prior for a $p \times p$ covariance matrix.

It reduces to the inverse gamma distribution if $p = 1$.

Say $\Sigma \sim \text{InvW}(\kappa, R)$ where $\kappa > p + 1$ and $R$ is a $p \times p$ covariance matrix are hyperparameters.

The PDF is

\[
f(\Sigma) \propto |\Sigma|^{-(\kappa+p+1)/2} \exp \left[ \frac{1}{2} \text{trace}(R\Sigma^{-1}) \right]
\]

The mean is $\frac{1}{\kappa-p-1} R$.
The hierarchical model is:

\[ Y_{ij} \sim \text{Normal}(\gamma_0i + X_i\gamma_1i, \sigma^2) \]

\[ \gamma_i \sim \text{Normal}(\beta, \Sigma) \]

\[ p(\beta) \propto 1 \]

\[ \sigma^2 \sim \text{InvGamma}(a, b) \]

\[ \Sigma \sim \text{InvWishart}(\kappa, R) \]

The full conditionals are all conjugate

**MCMC code:**

http://www4.stat.ncsu.edu/~reich/ABA/code/lmm.R

**JAGS:**

http://www4.stat.ncsu.edu/~reich/ABA/code/Jaw
Bone height data - fitted values

Subject 1

Subject 2

Subject 3

Subject 4

Subject 5

Subject 6

Population mean intercept

Population mean slope

Corr(\gamma_1, \gamma_2)

\beta_1

\beta_2

Correlation
Missing data models

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

- The model is

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

- 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

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

- 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 $\beta$?
Missing covariates

- Now say all responses are observed, but 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 \( \mathbf{X}_i = (X_{i1}, \ldots, X_{ip})^T \) 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 $\beta$ that properly accounts for uncertainty about the missing data.
Hierarchical linear regression model with missing data

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

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

▶ JAGS does not handle missing $X$ well

▶ Let’s use this as an opportunity to explore OpenBUGS
  http://www4.stat.ncsu.edu/~reich/ABA/code/Missing.R
Non- and Semi-parametric modeling

- Nonparametric (NP) methods attempt to analyze the data by making the fewest number of assumptions as possible.

- NP methods are generally more robust and flexible, but less powerful than correctly specified parametric models.

- Most frequentist NP methods completely avoid specifying a model.

- For example, a rank or sign test to compare two means.
Non- and Semi-parametric modeling

- Bayesian methods need a likelihood in order to obtain a posterior, so you can’t completely avoid specifying a model.

- Bayesian NP (BNP) then attempts to specify a model that is so flexible that it almost certainly captures the true model.

- One definition of the BNP model is one that has infinitely-many parameters.

- In some cases, NP models are difficult conceptually and computationally, and so semiparametric models with a large but finite number of parameters are useful approximations.
Parametric simple linear regression

Consider the classic parametric model:

\[ Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \quad \text{where} \quad \epsilon_i \sim N(0, \sigma^2). \]

Assumptions:

1. \( \epsilon_i \) are independent
2. \( \epsilon_i \) are Gaussian
3. The mean of \( Y_i \) is linear in \( X \).
4. The residual distribution does not depend on \( X \)

Alternatives:

1. Parametric alternatives such as a time series model.
2. Let \( \epsilon_i \sim F \), and place a prior on the distribution \( F \).
3. Let \( E(Y|X) = g(X) \) and put a prior on the function \( g \).
4. Heteroskedastic regression \( \text{Var}(\epsilon_i) = \exp(\alpha_0 + \alpha_1 X) \).

In 2-4 we are placing priors on functions, not parameters.
Nonparametric regression

- Let’s relax the assumption of linearity in the mean.

- The mean is \( g(X) \), where \( g \) is some function that relates \( X \) to \( E(Y|X) \).

- Parametric models include
  1. Linear: \( g(X) = \beta_0 + \beta_1 X \)
  2. Quadratic: \( g(X) = \beta_0 + \beta_1 X + \beta_2 X^2 \)
  3. Logistic: \( g(X) = \beta_0 + \beta_1 \frac{\exp[\beta_2 + \beta_3 X]}{1 + \exp[\beta_2 + \beta_3 X]} \).

- NP regression puts a prior on the curve \( g(X) \), rather than the parameters \( \beta_1, \ldots, \beta_p \) that determine the parametric model.
Semiparametric regression

- Semiparametric regression approximates the function $g$ using a finite basis expansion

$$
g(X) = \sum_{j=1}^{J} B_j(X) \beta_j
$$

where $B_j(X)$ are known basis functions and $\beta_j$ are unknown coefficients that determine the shape of $g$

- Example: the cubic spline basis functions are

$$
B_j(X) = (X - v_j)^3
$$

where $v_j$ are fixed knots that span the range of $X$

- Many other expansions exist: wavelets; Fourier, etc

- Fact: A basis expansion of $J$ terms can match the true curve $g$ at any $J$ points $X_1, ..., X_J$

- So increasing $J$ gives an arbitrarily flexible model
The model is \( Y_i \sim N(B_i^T \beta, \sigma^2) \), where \( \beta_j \sim N(0, \tau^2) \) and \( B_i \) is comprised of the known basis functions \( B_j(X_i) \).

Therefore, the model is usual linear regression model and is straightforward to fit using MCMC.

How to pick \( J \)?

Can we have more basis functions than observations?

What would you do if your prior was that \( g \) was probably quadratic, but you are not 100% sure about this. That is, your prior is that \( g(X) \approx \beta_0 + \beta_1 X + \beta_2 X^2 \).