CHAPTER 10 SLIDES START HERE
In hierarchical Bayesian estimation, we not only specify a prior on the data model’s parameter(s), but specify a further prior (called a hyperprior) for the hyperparameters.

This more complicated prior structure can be useful for modeling hierarchical data structures, also called multilevel data.

Multilevel data involves a hierarchy of nested populations, in which data could be measured for several levels of aggregation.

Examples:

- We could measure white-blood-cell counts for numerous patients within several hospitals.
- We could measure test scores for numerous students within several schools.
Hierarchical Bayes Estimation

- Assume we have data \( x \) from density \( f(x|\theta) \) with a parameter of interest \( \theta \).
- Typically we would choose a prior for \( \theta \) that depends on some hyperparameter(s) \( \psi \).
- Instead of choosing fixed values for \( \psi \), we could place a **hyperprior** \( p(\psi) \) on it.
- Note that this hierarchy could continue for any number of levels, but it is rare to need more than two levels for the prior structure.
Our posterior is then:

\[ \pi(\theta, \psi | x) \propto L(\theta | x)p(\theta | \psi)p(\psi) \]

Posterior inference about \( \theta \) is based on the \textit{marginal} posterior for \( \theta \):

\[ \pi(\theta | x) = \int_{\psi} \pi(\theta, \psi | x) d\psi \]

Except in simple situations, such analysis typically requires MCMC methods.
Example 1 (Economic data): Six economic indicators are measured at 44 timepoints \( x_1, \ldots, x_{44} \) (labeled 1, 2, \ldots, 44).

We model each indicator \( Y_i, i = 1, \ldots, 6 \) as a function of (centered) time as follows:

\[
Y_{ij} \sim N(\beta_{0i} + \beta_{1i}x_j, \tau)
\]
\[
\beta_{0i} \sim N(\mu_\beta, \tau_\beta_0)
\]
\[
\beta_{1i} \sim N(\mu_\beta_1, \tau_\beta_1)
\]
\[
\tau \sim \text{gamma}(0.01, 0.01)
\]
\[
\mu_\beta \sim N(0, 0.01), \quad \mu_\beta_1 \sim N(0, 0.01)
\]
\[
\tau_\beta_0 \sim \text{gamma}(0.01, 0.01), \quad \tau_\beta_1 \sim \text{gamma}(0.01, 0.01)
\]

See WinBUGS example for inference on \( \beta_{0i} \) and \( \beta_{1i} \), \( i = 1, 2, \ldots, 6 \).
Example 2 (Italian marriage data): Data are marriage counts (per 1000) in Italy for years from 1936 to 1951 (before, during, and after World War II).

We use a Poisson-Gamma hierarchical model that allows the Poisson mean to vary across years:

\[
Y_i \sim \text{Pois}(\lambda_i) \\
\lambda_i \sim \text{gamma}(\alpha, \beta) \\
\alpha \sim \text{gamma}(A, B) \\
\beta \sim \text{gamma}(C, D)
\]

and \(Y_1|\lambda_1, \ldots, Y_n|\lambda_n\) conditionally independent.

Note this allows the \(\lambda_i\)'s to be different, but following the same distribution.
Hierarchical Bayes Example 2

- It can be shown the full conditionals are:

  \[ \lambda_i | \alpha, \beta, y \sim \text{gamma}(y_i + \alpha, 1 + \beta) \]
  \[ \alpha | \beta, \lambda, y \sim \text{not a standard distribution} \]
  \[ \beta | \alpha, \lambda, y \sim \text{not a standard distribution} \]

- A Gibbs sampler can be implemented, e.g., in WinBUGS.
- The inference is on the \( \lambda_1, \ldots, \lambda_n \).
Recall for a fixed $n$, $X_1, X_2, \ldots, X_n$ are exchangeable if $p(X_1, \ldots, X_n) = p(X_{\pi_1}, \ldots, X_{\pi_n})$ for any permutation $(\pi_1, \ldots, \pi_n)$ of $(1, \ldots, n)$. (Finite exchangeability)

Infinite exchangeability implies that every finite subset of an infinite sequence $X_1, X_2, \ldots$ is exchangeable.

From de Finetti’s theorem: Exchangeable $\Rightarrow$ iid (True in infinite case; approximately true in finite case)
Consider **multilevel data**, where the observations come from, say, \( m \) groups:

- **Data:** \( Y_1, Y_2, \ldots, Y_m \) where each
  \[
  Y_j = [Y_{1j}, \ldots, Y_{nj}]' \quad \text{for } j = 1, \ldots, m.
  \]

We can often treat \( Y_{1j}, \ldots, Y_{nj} \) as exchangeable.

It then makes sense to treat the data in group \( j \) as **conditionally iid** given some group-specific parameter \( \theta_j \):

\[
Y_{1j}, \ldots, Y_{nj}|\theta_j \overset{iid}{\sim} p(y|\theta_j)
\]

Next, we can treat \( \theta_1, \ldots, \theta_m \) as exchangeable, if the groups are a random sample from a larger population of groups.

Again by de Finetti’s theorem:

\[
\theta_1, \ldots, \theta_m|\phi \overset{iid}{\sim} p(\theta|\phi)
\]
In this $m$-sample data analysis:

$p(y_{1j}, \ldots, y_{nj}|\theta_j)$ describes the within-group sampling variability

$p(\theta_1, \ldots, \theta_m|\phi)$ describes the between-group sampling variability

$p(\phi)$ describes uncertainty about $\phi$

We could continue the hierarchy, putting hyperpriors on the parameters in $p(\phi)$, but eventually we must stop.

The highest-level prior is often given a **diffuse** form.
Assume we have random samples from $m$ populations, having sample sizes $n_1, n_2, \ldots, n_m$.

We specify the hierarchical data model:

\[ Y_{1j}, \ldots, Y_{nj} | \mu_j, \sigma^2 \overset{iid}{\sim} N(\mu_j, \sigma^2) \quad \text{(within group-model)} \]

\[ \mu_j | \phi, \tau^2 \overset{iid}{\sim} N(\phi, \tau^2) \quad \text{(between-group model)} \]

This model assumes variability across group means, but group variances are assumed to be constant ($= \sigma^2$) across groups.
We place (independent) priors on the unknown parameters $\phi$, $\tau^2$ and $\sigma^2$:

\[
\frac{1}{\sigma^2} \sim \text{gamma}(\nu_1/2, \nu_1 \nu_2/2)
\]
\[
\frac{1}{\tau^2} \sim \text{gamma}(\eta_1/2, \eta_1 \eta_2/2)
\]
\[
\phi \sim N(\phi_0, \gamma^2)
\]
We must approximate the joint posterior

\[ \pi(\mu_1, \ldots, \mu_m, \phi, \tau^2, \sigma^2 | y_1, \ldots, y_m) \]

We will derive the full conditional for each parameter and use the Gibbs sampler to iteratively sample from these.

Note the joint posterior is

\[ \propto p(y_1, \ldots, y_m | \mu_1, \ldots, \mu_m, \phi, \tau^2, \sigma^2) \times p(\mu_1, \ldots, \mu_m | \phi, \tau^2, \sigma^2) p(\phi, \tau^2, \sigma^2) \]

\[ = \left[ \prod_{j=1}^{m} \prod_{i=1}^{n_j} p(y_{ij} | \mu_j, \sigma^2) \right] \left[ \prod_{j=1}^{m} p(\mu_j | \phi, \tau^2) \right] p(\phi)p(\tau^2)p(\sigma^2) \]

Note that conditional on \( \mu_j \) and \( \sigma^2 \), the joint density of the \( Y_{ij} \)'s does not depend on \( \phi \) and \( \tau^2 \).
From the above, we see the full conditionals for $\phi$ and $\tau^2$ satisfy:

\[ p(\phi | \mu_1, \ldots, \mu_m, \tau^2, \sigma^2, y_1, \ldots, y_m) \propto p(\phi) \prod_{j=1}^{m} p(\mu_j | \phi, \tau^2) \]

\[ p(\tau^2 | \mu_1, \ldots, \mu_m, \phi, \sigma^2, y_1, \ldots, y_m) \propto p(\tau^2) \prod_{j=1}^{m} p(\mu_j | \phi, \tau^2) \]
It can be shown that the full conditional for $\phi$ is normal and the full conditional for $\tau^2$ is inverse gamma. Specifically:

$$\phi | \mu_1, \ldots, \mu_m, \tau^2 \sim N \left( \frac{m \bar{\mu}}{\tau^2} + \frac{\phi_0}{\gamma^2}, \frac{1}{\frac{m}{\tau^2} + \frac{1}{\gamma^2}} \right)$$

and

$$\frac{1}{\tau^2} | \mu_1, \ldots, \mu_m, \phi \sim \text{gamma} \left( \frac{\eta_1 + m}{2}, \frac{\eta_1 \eta_2 + \sum_j (\mu_j - \phi)^2}{2} \right)$$

Similarly, the full conditional for any $\mu_j$ satisfies:

$$p(\mu_j | \phi, \tau^2, \sigma^2, y_1, \ldots, y_m) \propto p(\mu_j | \phi, \tau^2) \prod_{i=1}^{n_j} p(y_{ij} | \mu_j, \sigma^2)$$

Conditional on $\phi, \tau^2, \sigma^2, \mu_j$ is independent of the other $\mu$’s and of the data in the other groups.
Then it can be shown:

$$\mu_j | y_j, \sigma^2, \tau^2, \phi \sim N \left( \frac{n_j \bar{y}_j}{\sigma^2} + \frac{\phi}{\tau^2}, \frac{1}{n_j \sigma^2 + 1/\tau^2} \right)$$

Similarly, the full conditional for $\sigma^2$ is conditionally independent of $\{\phi, \tau^2\}$, given $\{y_1, \ldots, y_m, \mu_1, \ldots, \mu_m\}$:

$$p(\sigma^2 | \mu_1, \ldots, \mu_m, y_1, \ldots, y_m) \propto p(\sigma^2) \prod_{j=1}^m \prod_{i=1}^{n_j} p(y_{ij} | \mu_j, \sigma^2)$$

$$\propto (\sigma^2)^{-\nu_1/2+1} e^{-\nu_1 \nu_2 / 2\sigma^2} \left( \sigma^2 \right)^{-\frac{1}{2}} e^{-\frac{1}{2\sigma^2} \sum_j \sum_i (y_{ij} - \mu_j)^2}$$

Collecting terms, this is an inverse gamma, and:

$$\frac{1}{\sigma^2} | \mu, y_1, \ldots, y_m \sim \text{gamma} \left( \frac{1}{2} \left( \nu_1 + \sum_{j=1}^m n_j \right), \frac{1}{2} \left[ \nu_1 \nu_2 + \sum_j \sum_i (y_{ij} - \mu_j)^2 \right] \right)$$
Example 3 (Math scores): The data are math scores for 10th-grade students from $m = 100$ different urban high schools.

- The sample sizes $n_1, \ldots, n_m$ are quite different across schools.
- The nationwide total (between plus within) variance for this test is 100, and the nationwide mean is 50.

We choose the priors

$$\frac{1}{\sigma^2} \sim \text{gamma}(1/2, 100/2)$$

$$\frac{1}{\tau^2} \sim \text{gamma}(1/2, 100/2)$$

$$\phi \sim \mathcal{N}(50, 25)$$

We can then repeatedly cycle through $\phi^{[s]}, \tau^2[^{[s]}, \sigma^2[^{[s]}, \mu_1^{[s]}, \ldots, \mu_m^{[s]}$ (for $s = 1, \ldots, S$) using their full conditionals and the Gibbs sampler.

See R example with real schools data.