Chapter 5, Hierarchical Models

Aims

- Hierarchical models for combining information and meta-analysis
- Rat-tumor example
  - analysis with a fixed prior
  - analysis with historical data
- Exchangeability (de Finetti’s theorem)
- Posterior predictive distributions
- Fully Bayesian treatment of the hierarchical model
  - computation
  - Rat-tumor example (cont.)
Hierarchical or multi-stage models are a natural way to think about modeling information from partially exchangeable units.

- They may be tailored to model both the properties about the unit themselves

\[
y_i^s \mid \theta^s \sim f(y_i^s \mid \theta^s), \; i = 1, \ldots, n^s
\]

- and how these properties may among units

\[
\theta^s \mid \theta^* \sim g(\theta^*), \; s = 1, \ldots, S
\]

- along with a specification of prior distributions for the hyperparameters in the last stage

\[
\theta^* \sim h(\theta_0)
\]
Example

- Study of the effectiveness of cardiac treatments
- \( \theta^j \) is the survival probability for patients in hospital \( j \)
  
  *It might be reasonable to expect that estimates of the \( \theta^j \)'s, which represent a sample of hospitals, should be related each other* — \( \theta^s \sim g(\theta^*) \)
- \( \theta^* \) overall survival probability
  1. estimate \( \theta^j \)'s borrowing strength information from all the other hospitals — \( p(\theta^j \mid y^1, ..., y^J) \)
  2. estimate \( \theta^* \) taking into account the variability among hospitals—\( p(\theta^* \mid y^1, ..., y^J) \)
Hierarchical models for meta-analysis

- If there are several studies that address the same research question, one might be interested in combining the information from the individual studies in order to draw some sort of overall conclusion about the research question of interest.

- The studies may be thought of as belonging to a population of studies addressing the same research question, and the combining of individual studies in order to learn about the whole is referred to in the literature as *meta-analysis*. 
Analyzing a single experiment in the context of the historical data

- $\theta =$ probability of tumor in a population of female laboratory rats that receive zero dose of the drug
- $y/n =$ number of rats with tumor/number of rats
- current experiment $y/n = 4/14$

Analysis with a fixed prior distribution

\[
\begin{align*}
y \mid n, \theta & \sim Bin(n, \theta) \\
\theta & \sim Beta(\alpha, \beta) \\
\theta \mid y, n & \sim Beta(\alpha + 4, \beta + 10)
\end{align*}
\]
Analysis using historical data

- historical experiments \( y_j \mid n_j, j = 1, \ldots, 70 \)
- \( y_j \mid n_j, \theta_j \sim Bin(n_j, \theta_j) \)
- \( \theta_j \) is the study-specific mean
- \( \theta_j \sim Beta(\alpha, \beta) \)
- estimate \( \alpha \) and \( \beta \) with the observed mean and std dev of \( y_j/n_j, j = 1, \ldots, 70 \)

\[
0.136 = \frac{1}{70} \sum_{j=1}^{70} \frac{y_j}{n_j} = \hat{\mu} = \frac{\hat{\alpha}}{\hat{\alpha} + \hat{\beta}}
\]

\[
0.103 = \frac{1}{70} \sum_{j=1}^{70} \left( \frac{y_j}{n_j} - \hat{\mu} \right)^2 = \hat{\sigma}^2 = \frac{\hat{\alpha}\hat{\beta}}{(\hat{\alpha} + \hat{\beta})^2(\hat{\alpha} + \hat{\beta} + 1)}
\]

- \( \hat{\alpha} = 1.4, \hat{\beta} = 8.6 \)
- \( \theta_j \sim Beta(1.4, 8.6), y_{71} = 4, n_{71} - y_{71} = 10 \)
- \( \theta_j \mid y_{71}, n_{71}, \text{historical data} \sim Beta(5.4, 18.6) \)
Results of an "empirical Bayes approach"

- $E[\theta_{71} \mid \text{historical data}] = \frac{\hat{\alpha}}{\hat{\alpha} + \hat{\beta}} = 0.136$
- $\hat{\theta}_{71} = \frac{y_{71}}{n_{71}} = 0.286$
- $E[\theta_{71} \mid y_{71}, n_{71}, \text{historical data}] = 0.223$
- This is because the weight of the experience indicates that the number of tumors in the current experiment is unusually high.
- This is not a Bayesian approach because it is not based on any specified full probability model.
- The estimate of $\alpha$ and $\beta$ from $y$ is simply a starting point from which one can explore the idea of estimating the parameters of the population distribution.
Complete Bayesian Analysis
Define a probability model on the entire set of parameters and experiments and then perform a Bayesian analysis on the joint distribution of all model parameters

Empirical Bayes
Find point estimates of the population parameters using historical data. This approach can be viewed as approximation to the complete Bayesian analysis
Exchangeability

- $j = 1, \ldots, J$ is a set of experiments
- $y_1, \ldots, y_J$ data
- $\theta_1, \ldots, \theta_J$ experiment-specific parameters
- $p(y_1 \mid \theta_1, \ldots, y_J \mid \theta_J)$ specific likelihood
  $\theta_1, \ldots, \theta_J$ are exchangeable iff $p(\theta_1, \ldots, \theta_J)$ is invariant to permutations of indexes $j = 1, \ldots, J$

Ignorance $\rightarrow$ Exchangeability

*If no ordering or grouping of the parameters can be made, one must assume symmetry among the parameters in their prior distribution*
\( \theta_j \) is an independent draw from a population distribution governed by some unknown parameters \( \phi \)

\[
p(\theta \mid \phi) = \prod_{j=1}^{J} p(\theta_j \mid \phi)
\]

\[
p(\theta) = \int \left[ \prod_{j=1}^{J} p(\theta_j \mid \phi) \right] p(\phi) d\phi
\]

\( p(\theta) \) is a mixture of iid distributions

De Finetti’s Theorem

As \( J \to \infty \), any "suitable well-behaved" exchangeable distribution on \( \theta_1, \ldots, \theta_j \) can be written in the iid mixture form \( p(\theta) \)

Statistically, the iid mixture model characterizes parameters \( \theta \) from a common "superpopulation" that it is determined by the unknown hyperparameter \( \phi \)
Exchangeability when additional information is available on the units

- $x = (x_1, ..., x_J)$: experiment-specific covariates
- $\theta = (\theta_1, ..., \theta_J)$: experiment-specific parameters
- $\theta_1 | x_1, ..., \theta_J | x_J$: are exchangeable
- $p(\theta | x) = \int \left[ \prod_{j=1}^{J} p(\theta_j | \phi, x_j) \right] p(\phi | x) d\phi$
Objections to exchangeable models

- In rat tumor example, the 71 experiments were performed at different times, on different rats, and in different laboratories.
- Is it the exchangeable assumption acceptable?
- *that the experiments differ implies that the θ’s differ, but it might be perfectly acceptable to consider them as if drawn from a common distribution*
Full Bayesian treatment of the hierarchical model

- Posterior distribution of \((\phi, \theta)\)

\[
p(\phi, \theta) = p(\phi)p(\theta | \phi)
\]

\[
p(\phi, \theta | y) \propto p(\phi, \theta)p(y | \phi, \theta)
\]

\[
\propto p(\phi, \theta)p(y | \theta) \quad \text{(sometimes)}
\]
Posterior predictive distributions

- distribution of a failure observation $\tilde{y}$ corresponding to an existing $\theta_j$ (*additional rats from an existing experiment*)
- distribution of a failure observation $\tilde{y}$ corresponding to future $\theta_j$’s drawn from the same superpopulation (*results from a future experiment*)
- posterior predictive draws $\tilde{y}$ are based on the posterior draws $\tilde{\theta}$ for the existing experiment.
- posterior predictive draws $\tilde{y}$ are based on simulated $\tilde{\theta}$:
  - draw $\tilde{\phi}$ from $p(\phi \mid y)$
  - draw $\tilde{\theta}$ from $p(\theta \mid \tilde{\phi}, y)$
  - draw $\tilde{y}$ from $p(y \mid \tilde{\theta})$
**Computation with Hierarchical Models**

- \( \theta \) vector of parameters of interest
- \( \phi \) vector of nuisance parameters

AIM: to obtain simulations from the joint posterior distribution \( p(\theta, \phi \mid y) \) when the population distribution \( p(\theta, \phi) \) is conjugate to the likelihood \( p(y \mid \theta) \)

**Analytic derivation**

1. write \( p(\theta, \phi \mid y) \) in unnormalized form (immediate)
   \[
   p(\theta, \phi \mid y) \propto p(\phi)p(\theta \mid \phi)p(y \mid \theta)
   \]

2. determine analytically \( p(\theta \mid \phi, y) \)—(easy for conjugate normal model)

3. find \( p(\phi \mid y) = \int p(\theta, \phi \mid y)d\theta \)
Drawing simulations from the posterior distribution

1. Draw $\phi^* \sim p(\phi \mid y)$
2. Draw $\theta^* \sim p(\theta \mid \phi^*, y)$
3. If the factorization $p(\theta \mid \phi, y) = \prod p(\theta_j \mid \phi, y)$ holds, then the components $\theta_j$ can be drawn independently, one at a time
4. Draw $\tilde{y} \sim p(y \mid \theta^*)$
5. Repeat the steps $L$ times in order to obtain a set of $L$ draws
Full Bayesian analysis of rat tumors example

\( j = 1, \ldots, 71 \) experiments (Tarone, 1982)

\[
y_j \mid \theta_j, n_j \sim Bin(\theta_j, n_j)
\]

\[
\theta_j \mid \alpha, \beta \sim Beta(\alpha, \beta)
\]

\[
\alpha, \beta \sim \text{noninformative}
\]

1. \( p(\alpha, \beta, \theta \mid y) \propto p(\alpha, \beta)p(\theta \mid \alpha, \beta)p(y \mid \theta) \)
2. \( \theta_j \mid \alpha, \beta, y \sim Beta(\alpha + y_j, \beta + n_j - y_j) \)
3. simulate from \( p(\alpha, \beta \mid y) \)
   
   3.1 valuate \( p(\alpha, \beta \mid y) \) over a grid of points
   3.2 approximate it as a step-function
   3.3 sample \( \alpha^l \sim p(\alpha \mid y) \) and \( \beta^l \sim p(\beta \mid \alpha^l, y) \) using inverse cdf method
Setting up a noninformative prior distribution

Aim: found a diffuse hyperprior distribution for $\alpha$ and $\beta$
  - let $\gamma_1 = \log \frac{\alpha}{\beta}$ and $\delta_1 = \log(\alpha + \beta)$
  - $p_1(\gamma_1, \delta_1) \propto \text{constant}$. this prior leads an improper posterior
  - let $\gamma_2 = \frac{\alpha}{\alpha + \beta}$ and $\delta_2 = (\alpha + \beta)^{-1/2}$
  - $p_2(\gamma_2, \delta_2) \propto \text{constant}$, leads a proper posterior
  - The two priors are equivalent to:
    
    $p_1(\alpha, \beta) \propto (\alpha + \beta)^{-5/2}$
    $p_2(\alpha, \beta) \propto \alpha \beta (\alpha + \beta)^{-5/2}$
Computing the marginal posterior density

- Contour plot of the unnormalized marginal posterior \( p(\gamma, \delta \mid y) \)
- draw 1000 random samples from the joint posterior \( p(\alpha, \beta, \theta_1, \ldots, \theta_{71} \mid y) \) as follows:
  1. simulate \( \gamma^l, \delta^l \) from \( p(\gamma, \delta \mid y) \) with the discrete-grid sampling procedure
     1.1 for \( l = 1, \ldots, 1000 \)
     1.2 transform \( \gamma^l, \delta^l \rightarrow \alpha^l, \beta^l \)
  2. for each \( l \) draw \( \theta^l \sim Beta(\alpha^l + y_j, \beta^l + n_j - y_j) \)
  3. displays the results, for example histogram of \( ED50 = \alpha/\beta \)
Hierarchical Normal Model

- variance known (sec. 5.4)

\[
\begin{align*}
  y_{ij} & \sim N(\theta_j, \sigma^2), \quad i = 1, \ldots, n_j, j = 1, \ldots, J \\
  \theta_j & \sim N(\mu, \tau^2) \\
  \mu, \tau^2 & \sim p(\mu, \tau^2)
\end{align*}
\]

SAT coaching experiment

- variance unknown (sec. 9.8)

\[
\begin{align*}
  y_{ij} & \sim N(\theta_j, \sigma^2), \quad i = 1, \ldots, n_j, j = 1, \ldots, J \\
  \theta_j & \sim N(\mu, \tau^2) \\
  \mu, \sigma^2, \tau^2 & \sim p(\mu, \sigma^2, \tau^2)
\end{align*}
\]

DIET measurements

- approximating joint posterior distribution with direct simulation and with a Gibbs sampler
Classical random-effect analysis of the variance

\[ y_{ij} \sim N(\theta_j, \sigma^2) \]

\[ \bar{y}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} y_{ij} \sim N(\theta_j, \sigma_j^2), \quad \sigma_j^2 = \frac{\sigma^2}{n_j} \]

How do we estimate \( \theta_j \)? Two possibilities:

- Separate estimates: \( \bar{y}_j \) (bad choice when \( J = 20 \) and \( n_j = 2 \))
- Pooled estimates: \( \bar{y}_. = (\sum_j 1/\sigma_j^2 \bar{y}_j)/(\sum_j 1/\sigma_j^2) \)
- Weighted combination:

\[ \hat{\theta}_j = \lambda_j \bar{y}_j + (1 - \lambda_j) \bar{y}_., \quad 0 \leq \lambda_j \leq 1 \]

- keep in mind that \( \bar{y}_j \sim N(\theta_j, \sigma_j^2) \) can be a good approximation of the likelihood function even when \( y_{ij} \) are not normal.
Classical approach perform an analysis of variance $F$ test for differences among means

$$F = \frac{MS(\text{between})}{MS(\text{within})}$$

- if $F \gg 1$ we reject $H_0 : \tau^2 = 0$, and we use $\bar{y}_j$’s
- if $F \ll 1$ we accept $H_0 : \tau^2 = 0$, and we use $\bar{y}_.$
- $\bar{y}_j$ is the posterior mean of $\theta_j$ if $\theta_j \sim U(\infty, \infty)$
- $\bar{y}_.$ is the posterior mean of $\theta$ if $\theta_1 = \ldots = \theta_J = \theta$ and $\theta \sim U(\infty, \infty)$
- $\lambda_j \bar{y}_j + (1 - \lambda_j)\bar{y}_.$ is the posterior mean of $\theta_j$ if $\theta_j \sim N(\mu, \tau^2)$
- $\lambda_j = 1 \rightarrow \tau^2 = \infty$ separate
- $\lambda_j = 0 \rightarrow \tau^2 = 0$ pooling
Computations

- let $\theta = (\theta_1, ..., \theta_J)$
- joint prior $p(\mu, \tau^2) = p(\mu \mid \tau^2)p(\tau^2)$
- joint posterior
  \[
p(\theta, \mu, \tau \mid y) = p(\tau \mid y)p(\mu \mid \tau, y)p(\theta \mid \mu, \tau, y)
  \propto p(\tau \mid y)N(\hat{\mu}, V_\mu) \prod_{j=1}^J N(\hat{\theta}_j, V_j)
  \]
  where
  \[
  \hat{\mu} = \frac{\sum_j \frac{1}{\sigma_j^2 + \tau^2} \bar{y}_j}{\sum_j \frac{1}{\sigma_j^2 + \tau^2}}, \quad V_\mu^{-1} = \sum_j \frac{1}{\sigma_j^2 + \tau^2} \quad \alpha = \frac{1}{\sigma_j^2} \quad \beta = \frac{1}{\tau^2} \\
  \hat{\theta}_j = \frac{1}{\sigma_j^2} \bar{y}_j + \frac{1}{\tau^2} \mu \quad V_j^{-1} = \frac{1}{\sigma_j^2 + \frac{1}{\tau^2}} + \frac{1}{\tau^2} \]
- $p(\tau \mid y)$ is not available in closed form.
The Bayesian analysis under a normal hierarchical model provides a compromise that combines information from all the experiments without assuming all the $\theta_j$’s to be equal.

Hierarchical structuring of the models is an essential tool for achieving *partial pooling* of estimates, and compromising in a scientific way between alternative sources of information.
Table: Observed effects of special preparation on SAT-V scores in eight randomized experiments. Rubin (1981)

<table>
<thead>
<tr>
<th>School</th>
<th>treatment effect, $y_j$</th>
<th>error of effect estimate, $\sigma_j$</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>28.39</td>
<td>14.9</td>
</tr>
<tr>
<td>B</td>
<td>7.94</td>
<td>10.2</td>
</tr>
<tr>
<td>C</td>
<td>-2.75</td>
<td>16.3</td>
</tr>
<tr>
<td>D</td>
<td>6.82</td>
<td>11.0</td>
</tr>
<tr>
<td>E</td>
<td>-0.64</td>
<td>9.4</td>
</tr>
<tr>
<td>F</td>
<td>0.63</td>
<td>11.4</td>
</tr>
<tr>
<td>G</td>
<td>18.01</td>
<td>10.4</td>
</tr>
<tr>
<td>H</td>
<td>12.16</td>
<td>17.6</td>
</tr>
</tbody>
</table>

SAT coaching experiment

Separate randomized experiments were performed to estimate the effect of coaching programs for the SAT-V in each of 8 schools.
Separate estimates

- Looking at the table we see that it is difficult statistically to distinguish between any of the experiments.
- Treating each experiment separately, and applying the simple normal analysis, yield 95% posterior intervals that all overlap.

Difficulties

- \( \theta_A \mid y \sim N(28.4, 14.9^2) \)
- \( P(\theta_A > 28.4 \mid y_4) = \frac{1}{2} \) a doubtful statement, considering the results from the other 7 schools.
Pooled estimates

Under the hypothesis that all the experiments have the same effect and produce independent estimates of this common effect, we can suppose that $\theta_1 = \ldots = \theta_8$

$$E[\theta \mid y] = \left( \sum_j y_j / \sigma_j^2 \right) / \left( \sum_j 1 / \sigma_j^2 \right) = 7.9$$

$$\text{Var}[\theta \mid y] = \left( \sum_j 1 / \sigma_j^2 \right)^{-1} = 17.4$$

Difficulties

- $\theta_A \mid y \sim \mathcal{N}(7.9, 17.4^2)$
- $\theta_C \mid y \sim \mathcal{N}(7.9, 17.4^2)$
- $P(\theta_A - \theta_C < 0 \mid y) = \frac{1}{2}$ it is difficult to justify from the table
Hierarchical Model with known variance

\[ y_j \sim N(\theta_j, \sigma_j^2) \]
\[ \theta_j \sim N(\mu, \tau^2) \]
\[ \mu \sim N(0, 100) \]
\[ \tau^2 \sim IG(AA, BB) \]

Two options

- Direct sampling

\[ p(\tau, \mu, \theta | y) \propto p(\tau | y)p(\mu | \tau, y) \prod_{j=1}^{J} p(\theta_j | \mu, \tau, y) \]

- Gibbs sampling: write down

\[ p(\theta_j | \mu, \tau, y) = ? \]
\[ p(\mu | \theta_1, ..., \theta_J, \tau, y) = ? \]
Table: Coagulation time in seconds for blood drawn from 24 animals randomly allocated to 4 different diets. Different treatments have different numbers of observations because the randomization was unrestricted (Box, Hunter (1978))

<table>
<thead>
<tr>
<th>Diet</th>
<th>Measurement</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>62,60,63,59</td>
</tr>
<tr>
<td>B</td>
<td>63,67,71,64,65,66</td>
</tr>
<tr>
<td>C</td>
<td>68,66,71,67,68,68</td>
</tr>
<tr>
<td>D</td>
<td>56,62,60,61,63,64,63,59</td>
</tr>
</tbody>
</table>

DIET-measurements example
Hierarchical Normal Model with unknown variance

\[ y_{ij} \sim N(\theta_j, \sigma^2), \ i = 1, \ldots, n_j \]
\[ \theta_j \sim N(\mu, \tau^2), \ j = 1, \ldots, J, \ n = \sum n_j \]
\[ \mu \sim N(0, 100), \sigma^2 \sim IG(AA1, BB1), \tau^2 \sim IG(AA2, BB2) \]

- if we were assign a uniform prior to log \( \tau \), the posterior would be improper (sec. 5)
- crude estimates

\[ \hat{\theta}_j = \bar{y}_j \]
\[ \hat{\sigma}^2 = \sum_j \frac{1}{n_j - 1} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2 \]
\[ \hat{\mu} = \frac{1}{J} \sum_j \hat{\theta}_j, \hat{\tau}^2 = \frac{1}{J - 1} \sum_j (\hat{\theta}_j - \hat{\mu})^2 \]
Implement a Gibbs sampler

List of the full conditional distributions

\[ \theta_j \mid y, \mu, \sigma, \tau \sim N\left( \frac{\mu}{\tau^2} + \frac{n_j \bar{y}_j}{\sigma^2} \left( \frac{1}{\tau^2} + \frac{n_j}{\sigma^2} \right)^{-1}, \left( \frac{1}{\tau^2} + \frac{n_j}{\sigma^2} \right)^{-1} \right) \]

\[ \mu \mid y, \theta_j, \sigma, \tau \sim N\left( \frac{1}{J} \sum \theta_j, \frac{\tau^2}{J} \right) \]

\[ \sigma^2 \mid y, \theta_1, \ldots, \theta_J, \mu, \tau \sim IG\left(AA1 + \frac{n}{2}, \frac{1}{2} \sum_{ij} (y_{ij} - \theta_j)^2 \right) \]

\[ \tau^2 \mid y, \theta_1, \ldots, \theta_J, \mu, \sigma \sim IG\left(AA2 + \frac{J}{2}, \frac{1}{2} \sum_j (\theta_j - \mu)^2 \right) \]