Hierarchical Modeling with Longitudinal (Panel) Data

Gibbs Sampling in Hierarchical Models

Econ 690

Purdue University

March 19, 2012
In many models in economics and statistics, it is natural to introduce some kind of structure relating the parameters of the model.

For example, one might wish to express some degree of “similarity” across parameters of a model by assuming that they are drawn from a common population distribution.

The parameters of the population distribution are also of interest.

In Bayesian terms, such specifications can be accommodated by the appropriate choice of priors on the model parameters, while in Frequentist parlance, these are termed “random coefficient” models.
Consider the following most basic version of a longitudinal (panel) data model:

\[ y_{it} \] refers to the outcomes for individual (or more generally, group) \( i \) at time \( t \), and \( \alpha_i \) is a person (or group) specific random effect.

We assume \( i = 1, 2, \ldots, N \) and \( t = 1, 2, \ldots, T \) (i.e., a balanced panel).
For this model, we will do the following:

1. (a) Comment on how the presence of the random effects $\alpha_i$ accounts for correlation patterns within individuals over time.

2. (b) Derive the conditional posterior distribution $p(\alpha_i | \alpha, \sigma_\epsilon^2, \sigma_\alpha^2, y)$.

3. (c) Obtain the mean of the conditional posterior distribution in (b). Comment on its relationship to a shrinkage estimator. (These are estimators that are typically written as some sort of weighted average of a “data” term and a prior term). How does the mean change as $T$ and $\sigma_\epsilon^2 / \sigma_\alpha^2$ change?
(a) Conditional on the random effects \( \{\alpha_i\}_{i=1}^N \), the \( y_{it} \) are independent.

However, marginalized over the random effects, outcomes are correlated within individuals over time.

To see this, note that we can write our model equivalently as:

\[
y_{it} = \alpha + u_i + \epsilon_{it},
\]

where we have rewritten our “random effect” specification as

\[
\alpha_i = \alpha + u_i, \quad u_i \stackrel{iid}{\sim} N(0, \sigma^2_\alpha)
\]
Thus for \( t \neq s \),

\[
\text{Cov}(y_{it}, y_{is} | \alpha, \sigma^2_\epsilon, \sigma^2_\alpha) = \text{Cov}(u_i + \epsilon_{it}, u_i + \epsilon_{is}) = \text{Cov}(u_i, u_i) = \text{Var}(u_i) = \sigma^2_\alpha,
\]

so that outcomes are correlated over time \textit{within individuals}.

However, the model does not permit any degree of correlation between the outcomes of different individuals.
From previous results relating to the derivation of conditional posterior distributions for regression parameters in a linear model, we can obtain:

where

and

with $\iota_T$ denoting a $T \times 1$ vector of ones.
The mean of this conditional posterior distribution is easily obtained from our solution in (b):

Let

\[ w = w(T, [\sigma^2_\epsilon / \sigma^2_\alpha]) \equiv \frac{T}{T + (\sigma^2_\epsilon / \sigma^2_\alpha)}. \]
We can then write

\[ E(\alpha_i | \beta, \sigma^2_\epsilon, \sigma^2_\alpha, y) = w \bar{y}_i + (1 - w) \alpha. \]

This is in the form of a **shrinkage estimator**, where the conditional posterior mean of \( \alpha_i \) is a weighted average of the averaged outcomes for individual \( i \), \( \bar{y}_i \), and the common mean for all individuals, \( \alpha \).

As \( T \to 1 \), the weight \( w \) places all mass on \( \bar{y}_i \).

On the other hand, if \( \sigma^2_\epsilon \) is large relative to \( \sigma^2_\alpha \), (and \( T \) is small or moderate), the common mean \( \alpha \) will get substantial weight.

The “fixed effect” formulation of this model is often criticized for overfitting.
We illustrate the use of the Gibbs sampler in such models with the celebrated “rat growth dataset” of Gelfand et al (1990).

30 different rats are weighed at 5 different points in time.

We denote the weight of rat $i$ at measurement $j$ as $y_{ij}$ and let $x_{ij}$ denote the age of the $i^{th}$ rat at the $j^{th}$ measurement.

Since each of the rats were weighed at exactly the same number of days since birth, we have

$$x_{i1} = 8, \ x_{i2} = 15, \ x_{i3} = 22, \ x_{i4} = 29, \ x_{i5} = 36 \ \forall i.$$ 

The rat growth data set is provided on the following page:
Rat Growth Data from Gelfand et al (1990).

<table>
<thead>
<tr>
<th>Rat</th>
<th>Weight Measurements</th>
<th>Rat</th>
<th>Weight Measurements</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$y_{i1}$</td>
<td>$y_{i2}$</td>
<td>$y_{i3}$</td>
</tr>
<tr>
<td>1</td>
<td>151</td>
<td>199</td>
<td>246</td>
</tr>
<tr>
<td>2</td>
<td>145</td>
<td>199</td>
<td>249</td>
</tr>
<tr>
<td>3</td>
<td>147</td>
<td>214</td>
<td>263</td>
</tr>
<tr>
<td>4</td>
<td>155</td>
<td>200</td>
<td>237</td>
</tr>
<tr>
<td>5</td>
<td>135</td>
<td>188</td>
<td>230</td>
</tr>
<tr>
<td>6</td>
<td>159</td>
<td>210</td>
<td>252</td>
</tr>
<tr>
<td>7</td>
<td>141</td>
<td>189</td>
<td>231</td>
</tr>
<tr>
<td>8</td>
<td>159</td>
<td>201</td>
<td>248</td>
</tr>
<tr>
<td>9</td>
<td>177</td>
<td>236</td>
<td>285</td>
</tr>
<tr>
<td>10</td>
<td>134</td>
<td>182</td>
<td>220</td>
</tr>
<tr>
<td>11</td>
<td>160</td>
<td>208</td>
<td>261</td>
</tr>
<tr>
<td>12</td>
<td>143</td>
<td>188</td>
<td>220</td>
</tr>
<tr>
<td>13</td>
<td>154</td>
<td>200</td>
<td>244</td>
</tr>
<tr>
<td>14</td>
<td>171</td>
<td>221</td>
<td>270</td>
</tr>
<tr>
<td>15</td>
<td>163</td>
<td>216</td>
<td>242</td>
</tr>
</tbody>
</table>
In our model, we want to permit unit-specific variation in birth and growth rates. This leads us to specify the following model:

so that each rat possesses its own intercept $\alpha_i$ and growth rate $\beta_i$. 
We also assume that the rats share some degree of “similarity” in their weight at birth and rates of growth.

Thus, we assume that the intercept and slope parameters are drawn from a common Normal population:

\[ \alpha_0 = \theta_0(1) \]

\[ \beta_0 = \theta_0(2) \]

We interpret \( \alpha_0 = \theta_0(1) \) as the population average weight at birth and \( \beta_0 = \theta_0(2) \) as the population average growth rate.

The diagonal elements of \( \Sigma \) quantify the variation around these population means. (What would we expect the sign of \( \Sigma \)’s off-diagonal to be)?
We complete our Bayesian analysis by specifying the following priors:

\[ \sigma^2|a, b \sim IG(a, b) \]
\[ \theta_0|\eta, C \sim N(\eta, C) \]
\[ \Sigma^{-1}|\rho, R \sim W([\rho R]^{-1}, \rho), \]

with \( W \) denoting the Wishart distribution. We now seek to describe how the Gibbs sampler can be employed to fit this hierarchical model.
Given the assumed conditional independence across observations, the joint posterior distribution for all the parameters of this model can be written as:

\[
p(\Gamma | y) \propto \prod_{i=1}^{30} p(y_i | x_i, \theta_i, \sigma^2) p(\theta_i | \theta_0, \Sigma^{-1}) \] \[\times p(\theta_0 | \eta, C)p(\sigma^2 | a, b)p(\Sigma^{-1} | \rho, R),
\]

where \( \Gamma \equiv \{\theta_i\}, \theta_0, \Sigma^{-1}, \sigma^2 \) denotes all the parameters of the model. We have stacked the observations over time for each individual rat so that

\[
y_i = \begin{bmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{i5} \end{bmatrix} \quad \text{and} \quad X_i = \begin{bmatrix} 1 & x_{i1} \\ 1 & x_{i2} \\ \vdots & \vdots \\ 1 & x_{i5} \end{bmatrix}.
\]
Fitting this model via the Gibbs sampler requires the derivation of four posterior conditional distributions:

1. $p(\theta_i | \Gamma_{-\theta_i}, y)$.

2. $p(\theta_0 | \Gamma_{-\theta_0}, y)$.

3. $p(\sigma^2 | \Gamma_{-\sigma^2}, y)$.

4. $p(\Sigma^{-1} | \Gamma_{-\Sigma^{-1}}, y)$.

We will derive each of these densities.
As for the complete posterior conditional for $\theta_i$, we note:

This fits directly into our standard linear regression result, applying Lindley and Smith (1972):

where
As for the posterior conditional for $\theta_0$, we first obtain

$$
\begin{bmatrix}
\theta_1 \\
\theta_2 \\
\vdots \\
\theta_{30}
\end{bmatrix} =
\begin{bmatrix}
l_2 \\
l_2 \\
\vdots \\
l_2
\end{bmatrix}
\theta_0 +
\begin{bmatrix}
u_1 \\
u_2 \\
\vdots \\
u_{30}
\end{bmatrix},
$$

Since the second stage of our model specifies $p(\theta_i|\theta_0, \Sigma^{-1})$ as iid, we can write
Equivalently, we can write:

\[ \tilde{\theta} = \tilde{I} \theta_0 + \tilde{u}, \]

with \( \tilde{\theta} = [\theta'_1 \theta'_2 \cdots \theta'_{30}]' \), \( \tilde{I} = [I_2 I_2 \cdots I_2]' \), \( \tilde{u} = [u'_1 u'_2 \cdots u'_{30}]' \) and \( E(\tilde{u} \tilde{u}') = I_{30} \otimes \Sigma \).

In this form, we can again apply our well-known result to obtain:

\[ \theta_0 | \Gamma_{-\theta_0}, y \sim N(D_{\theta_0} d_{\theta_0}, D_{\theta_0}) \]

where

\[
D_{\theta_0} = \left( \tilde{I}' (I_{30} \otimes \Sigma^{-1}) \tilde{I} + C^{-1} \right)^{-1} = (30\Sigma^{-1} + C^{-1})^{-1}
\]

\[
d_{\theta_0} = (\tilde{I}' (I_{30} \otimes \Sigma^{-1}) \tilde{\theta} + C^{-1} \eta) = (30\Sigma^{-1} \tilde{\theta} + C^{-1} \eta),
\]

where \( \tilde{\theta} = (1/30) \sum_{i=1}^{30} \theta_i \).
As for the posterior conditional for $\sigma^2$, we obtain

Thus,

where $N = 5(30) = 150$. 
Finally, for the posterior conditional for $\Sigma^{-1}$, we obtain

Therefore,
We fit this model using priors of the forms:

\[ \eta = \begin{bmatrix} 100 \\ 15 \end{bmatrix}, \quad C = \begin{bmatrix} 40^2 & 0 \\ 0 & 10^2 \end{bmatrix}, \quad \rho = 5, \quad R = \begin{bmatrix} 10^2 & 0 \\ 0 & .5^2 \end{bmatrix}, \]

\[ a = 3, \quad b = 1/40. \]

The sampler was run for 10,000 iterations, and the first 500 were discarded as the burn-in.

In the next two graphs, we provide some suggestive evidence of rapid convergence, and also that the chain tends to mix reasonably well.
Hierarchical Modeling with Longitudinal (Panel) Data

Justin L. Tobias

Hierarchical Models
<table>
<thead>
<tr>
<th>Parameter</th>
<th>Lag 1</th>
<th>Lag 5</th>
<th>Lag 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha_0$</td>
<td>.24</td>
<td>.010</td>
<td>.007</td>
</tr>
<tr>
<td>$\beta_0$</td>
<td>.22</td>
<td>-.004</td>
<td>-.009</td>
</tr>
<tr>
<td>$\sigma^2_\alpha$</td>
<td>.36</td>
<td>.020</td>
<td>-.010</td>
</tr>
<tr>
<td>$\rho_{\alpha,\beta}$</td>
<td>.37</td>
<td>.036</td>
<td>.003</td>
</tr>
<tr>
<td>$\alpha_{15}$</td>
<td>.18</td>
<td>.025</td>
<td>.018</td>
</tr>
</tbody>
</table>
### Table 12.2: Posterior Quantities for a Selection of Parameters

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Post Mean</th>
<th>Post Std.</th>
<th>10\textsuperscript{th} Percentile</th>
<th>90\textsuperscript{th} Percentile</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha_0$</td>
<td>106.6</td>
<td>2.34</td>
<td>103.7</td>
<td>109.5</td>
</tr>
<tr>
<td>$\beta_0$</td>
<td>6.18</td>
<td>.106</td>
<td>6.05</td>
<td>6.31</td>
</tr>
<tr>
<td>$\sigma^2_\alpha$</td>
<td>124.5</td>
<td>42.41</td>
<td>77.03</td>
<td>179.52</td>
</tr>
<tr>
<td>$\sigma^2_\beta$</td>
<td>.275</td>
<td>.088</td>
<td>.179</td>
<td>.389</td>
</tr>
<tr>
<td>$\rho_{\alpha,\beta}$</td>
<td>-.120</td>
<td>.211</td>
<td>-.390</td>
<td>.161</td>
</tr>
<tr>
<td>$\alpha_{10}$</td>
<td>93.76</td>
<td>5.24</td>
<td>87.09</td>
<td>100.59</td>
</tr>
<tr>
<td>$\alpha_{25}$</td>
<td>86.91</td>
<td>5.81</td>
<td>79.51</td>
<td>94.45</td>
</tr>
<tr>
<td>$\beta_{10}$</td>
<td>5.70</td>
<td>.217</td>
<td>5.43</td>
<td>5.98</td>
</tr>
<tr>
<td>$\beta_{25}$</td>
<td>6.75</td>
<td>.243</td>
<td>6.44</td>
<td>7.05</td>
</tr>
</tbody>
</table>
Hierarchical Modeling with Longitudinal (Panel) Data

- Intercepts
  - 5
  - 5.5
  - 6
  - 6.5
  - 7
  - 7.5

- Growth Rates
  - OLS Estimates
  - Hierarchical Model

- OLS Estimates
- Hierarchical Model

Justin L. Tobias
Hierarchical Models
A second application follows (Krueger 1998; Krueger and Whitmore 2001) and applies our model to analyze data from Project STAR (Student/Teacher Achievement Ratio).

Project STAR was an experiment in Tennessee that randomly assigned students to one of three types of classes - small class, regular size class, and regular size class with a teacher’s aide (regular/aide class).

The dependent variable is the average of a reading percentile score and math percentile score of a Project STAR student.

There are two explanatory variables - a dummy variable indicating whether a student is assigned to a small class and another indicating assignment to a regular/aide class. The default category, therefore, is assignment to regular class.
The Project STAR data we use contains 79 participating schools with a total of 5,726 students who entered the project during kindergarten.

We focus on the achievement measure taken at the end of the kindergarten year and consider heterogeneity of treatment impacts across schools.

Therefore, in this application of the model in (47), \( i \) denotes the school and \( j/\ t \) no longer represents a time index but, instead, denotes the student within a school.
Table 4: Posterior means, standard deviations and probabilities of being positive of the parameters

| Parameter                                             | $E(\beta|D)$ | $\text{Std}(\beta|D)$ | $\text{Pr}(\beta > 0|D)$ |
|-------------------------------------------------------|--------------|------------------------|--------------------------|
| $\beta_0$ (intercept)                                 | 51           | 1.82                   | 1                        |
| $\beta_1$ (small class)                               | 5.48         | 1.44                   | 1                        |
| $\beta_2$ (regular/aide class)                        | 0.311        | 1.26                   | 0.596                    |
| $\sqrt{\sigma^2}$                                    | 22.9         | 0.221                  | 1                        |
| $\sqrt{\Sigma_{\beta}(1, 1)}$                        | 15.2         | 1.32                   | 1                        |
| $\sqrt{\Sigma_{\beta}(2, 2)}$                        | 10.6         | 1.24                   | 1                        |
| $\sqrt{\Sigma_{\beta}(3, 3)}$                        | 8.93         | 1.14                   | 1                        |
| $\Sigma_{\beta}(1, 2)/\sqrt{\Sigma_{\beta}(1, 1) \times \Sigma_{\beta}(2, 2)}$ | -0.454      | 0.111                  | 0.000125                 |
| $\Sigma_{\beta}(1, 3)/\sqrt{\Sigma_{\beta}(1, 1) \times \Sigma_{\beta}(3, 3)}$ | -0.483      | 0.111                  | 0.000125                 |
| $\Sigma_{\beta}(2, 3)/\sqrt{\Sigma_{\beta}(2, 2) \times \Sigma_{\beta}(3, 3)}$ | 0.548       | 0.118                  | 1                        |
Evidence of small class size effect, no strong evidence of aide effect.

Evidence of heterogeneity of impacts across schools.

Strong correlation among school-level parameters ... what is the/a interpretation for this result?
Let us now turn to a restricted (and perhaps more widely-used) version of our previous model.

Note that, in the specification just presented, both the intercept and slope (or, more generally, set of slopes), were permitted to vary across the units.

In a restricted version of this model, perhaps the “default” panel data specification in economics, we may permit the intercept to vary across individuals, but restrict the other regression coefficients to be constant across individuals.

Such a model is termed a **mixed model**.
Formally, we consider a specification of the form:

\[ y_{it} = \alpha_i + x_{it}\beta + \epsilon_{it}, \quad \epsilon_{it} \overset{iid}{\sim} N(0, \sigma^2_{\epsilon}) \]

\[ \alpha_i \overset{iid}{\sim} N(\alpha, \sigma^2_{\alpha}). \]

For this model we employ independent priors of the form:

\[ \beta \sim N(\beta_0, V_{\beta}) \]

\[ \alpha \sim N(\alpha_0, V_{\alpha}) \]

\[ \sigma^2_{\epsilon} \sim IG(e_1, e_2) \]

\[ \sigma^2_{\alpha} \sim IG(a_1, a_2). \]
We seek to do the following:

- (a) Derive the complete posterior conditionals
  
  \[ p(\alpha_i | \beta, \alpha, \sigma_\epsilon^2, \sigma_\alpha^2, y), \quad \text{and} \quad p(\beta | \{\alpha_i\}, \alpha, \sigma_\epsilon^2, \sigma_\alpha^2, y), \]

- (b) Describe how one could use a *blocking* or *grouping* step [e.g., Chib and Carlin (1999)] to obtain draws directly from the *joint posterior conditional* \( p(\{\alpha_i\}, \beta | \alpha, \sigma_\epsilon^2, \sigma_\alpha^2, y) \).

- (c) Describe how the Gibbs sampler can be used to fit the model, given your result in (b). Would you expect any improvements in this blocked algorithm relative to the standard Gibbs algorithm in (a)?

- (d) How does your answer in (c) change for the case of an *unbalanced panel* where \( T = T_i \)?
As for (a), the complete posterior conditionals can be obtained in a straightforward manner. Specifically, we obtain

\[ p(\alpha_i | \beta, \alpha, \sigma_e^2, \sigma_\alpha^2, y) \sim N(Dd, D) \]

where

\[
D = \left( T/\sigma_e^2 + 1/\sigma_\alpha^2 \right)^{-1}, \quad d = \sum_{t=1}^{T} \frac{(y_{it} - x_{it} \beta)/\sigma_e^2 + \alpha/\sigma_\alpha^2}{\sigma_e^2 + \alpha/\sigma_\alpha^2}.
\]
The complete posterior conditional for $\beta$ follows similarly:

$$
\beta | \{ \alpha_i \}, \alpha, \sigma^2_\epsilon, \sigma^2_\alpha, y \sim N(Hh, H),
$$

where

$$
H = \left( X'X/\sigma^2_\epsilon + V_\beta^{-1} \right)^{-1}, \quad h = X'(y - \overline{\alpha})/\sigma^2_\epsilon + V_\beta^{-1}\beta_0,
$$

with

$$
\overline{\alpha} = [(\iota^T\alpha_1)' (\iota^T\alpha_2)' \cdots (\iota^T\alpha_N)']'.
$$
(b) Instead of the strategy described in (a), we seek to draw the random effects \( \{\alpha_i\} \) and “fixed effects” \( \beta \) in a single block. This strategy of grouping together correlated parameters will generally facilitate the mixing of the chain and thereby reduce numerical standard errors associated with the Gibbs sampling estimates.

We break this joint posterior conditional into the following two pieces:
The assumptions of our model imply that the random effects \( \{ \alpha_i \} \) are conditionally independent, so that

\[
p(\{\alpha_i\}, \beta|\alpha, \sigma^2_{\epsilon}, \sigma^2_{\alpha}, y) = \prod_{i=1}^{N} p(\alpha_i|\beta, \alpha, \sigma^2_{\epsilon}, \sigma^2_{\alpha}, y) p(\beta|\alpha, \sigma^2_{\epsilon}, \sigma^2_{\alpha}, y).
\]

This suggests that one can draw from this joint posterior conditional via the method of composition by first drawing from \( p(\beta|\alpha, \sigma^2_{\epsilon}, \sigma^2_{\alpha}, y) \) and then drawing each \( \alpha_i \) independently from its complete posterior conditional distribution.
We now derive the conditional posterior distribution for $\beta$, marginalized over the random effects. Note that our model can be rewritten as follows:

$$u_i \sim iid \sim N(0, \sigma^2_\alpha).$$

Let $v_{it} = u_i + \epsilon_{it}$. If we stack this equation over $t$ within $i$ we obtain:

$$y_i = [y_{i1} \ y_{i2} \ \cdots \ y_{iT}]', \quad x_i = [x'_{i1} \ x'_{i2} \ \cdots \ x'_{iT}]', \text{ and}$$

$$v_i = [v_{i1} \ v_{i2} \ \cdots \ v_{iT}]'.$$
Stacking again over \( i \) we obtain:

\[
y = i_{NT} \alpha + X \beta + \nu,
\]

where

\[
E(\nu \nu') = I_N \otimes \Sigma.
\]
In this form, we can now appeal to our standard results for the regression model to obtain

$$
\beta|\alpha, \sigma^2_{\epsilon}, \sigma^2_{\alpha}, y \sim N(Gg, G)
$$

where

$$
G = \left(X'(I_N \otimes \Sigma^{-1})X + V_{\beta}^{-1}\right)^{-1} = \left(\sum_{i=1}^{N} x_i'\Sigma^{-1}x_i + V_{\beta}^{-1}\right)^{-1}
$$

and

$$
g = X'(I_N \otimes \Sigma^{-1})(y - \iota_{NT}\alpha) + V_{\beta}^{-1}\beta_0 = \sum_{i=1}^{N} x_i'\Sigma^{-1}(y_i - \iota_{T}\alpha) + V_{\beta}^{-1}\beta_0.
$$

Thus, to sample from the desired joint conditional, you first sample $\beta$ from the distribution given above and then sample the random effects independently from their complete conditional posterior distributions.
Finally, it is also worth noting that $\beta$ and $\alpha$ could be drawn together in the first step of this process.

That is, one could draw from the joint posterior conditional

$$p(\{\alpha_i\}, \alpha, \beta | \sigma^2_\epsilon, \sigma^2_\alpha, y) = \left[ \prod_{i=1}^{N} p(\{\alpha_i\} | \beta, \alpha, \sigma^2_\alpha, \sigma^2_\epsilon, y) \right] p(\beta, \alpha | \sigma^2_\epsilon, \sigma^2_\alpha, y)$$

in a similar way as described above.

We would expect the mixing of such chains to improve relative to the standard (or unblocked) Gibbs sampler. In the limiting case, where we can sample in a single block, we are back to iid sampling!
(c) Given the result in (b), we now need to obtain the remaining complete conditionals. These are given as follows:

\[ \alpha | \{ \alpha_i \}, \beta, \sigma^2_\epsilon, \sigma^2_\alpha, y \sim N(Rr, R) \]

where

\[ R = \left( \frac{N}{\sigma^2_\alpha + V^{-1}_\alpha} \right)^{-1}, \quad r = \sum_{i=1}^{N} \frac{\alpha_i}{\sigma^2_\alpha + V^{-1}_\alpha} \alpha_0. \]

\[ \sigma^2_\alpha | \{ \alpha_i \}, \beta, \sigma^2_\epsilon, y \sim IG \left( (N/2) + a_1, \left[ a^{-1}_2 + 0.5 \sum_{i=1}^{N} (\alpha_i - \alpha)^2 \right]^{-1} \right). \]

\[ \sigma^2_\epsilon | \{ \alpha_i \}, \beta, \sigma^2_\alpha, y \sim IG \left( (NT/2) + e_1, \left[ e^{-1}_2 + 0.5(y - \iota_{NT}\alpha - X\beta)'(y - \iota_{NT}\alpha - X\beta) \right]^{-1} \right). \]
Only slight changes are required in the case of unbalanced panels.

Let $NT$ continue to denote the total number of observations, $NT \equiv \sum_{i=1}^{N} T_i$.

In addition, let $\Sigma_i \equiv \sigma^2_\epsilon I_{T_i} + \sigma^2_\alpha \epsilon T_i \epsilon'_{T_i}$.

Replacing $\Sigma$ with $\Sigma_i$ and $T$ with $T_i$, as appropriate in the above formulae, is all that is required.
Further Reading

Illustration of Bayesian inference in normal data models using Gibbs sampling.
*JASA* 85, 972-985.

On MCMC Sampling in hierarchical longitudinal models.