# Hierarchical Models for Heterogeneous Preferences

*(KPT Ch. 7)*

AAEC 6564
Instructor: KLAUS MOELTNER

Matlab scripts: `mod6_data, mod6_HNRM_v1, mod6_HNRM_v2, mod6_HNRM_outage, mod6_HNRM_outage_chib, mod6_HNRM_outage_nocovs, mod6_HNRM_outage_chib, mod6_HNRM_predict`

Matlab functions: `gs_HNRM_v1, gs_HNRM_v2, gs_HNRM_chib, gs_HNRM_nocovs, gs_HNRM_nocovs_chib`

"Hierarchical modeling" has become quite popular in both Classical and Bayesian analysis. These models have one or more additional distributional layers between the base structure and the priors for some of the parameters. Hierarchical models (HMs) can be estimated via a straightforward extension of the models we have already discussed. In this Session we will consider a Hierarchical Regression Model (HRM) with some, but not all, parameters taking such a "hyper-distribution"'. This is often called a regression model with "mixed coefficients", or a "*mixed effects*"-model.

## Hierarchical normal regression model

HMs require multiple observations per individual (or other observational unit of interest). Thus, panel data models with multiple observations over time per person are a special case of HMs. Any other grouping leading to multiple individual observations qualifies as well, such as visits to several hiking trails per individual, demand for several commodities per household, etc. For simplicity let's assume we have the same number of observations (say *J*) per person, for a total of *N* individuals, although such a "balanced" setup is not a requirement. Thus the total sample size is $n = NJ$. The model at the individual level is then given as

$$\begin{aligned} &\underset{Jx1}{\mathbf{y_i}} = \underset{k_f x1}{\mathbf{X_{fi}}\,\boldsymbol{\beta}_\mathbf{f}} + \underset{k_r x1}{\mathbf{X_{ri}}\,\boldsymbol{\beta}_\mathbf{ri}} + \boldsymbol{\varepsilon}_\mathbf{i} \qquad \boldsymbol{\varepsilon}_\mathbf{i} \sim n\left(0, \sigma^2 \mathbf{I_J}\right) \\ &\boldsymbol{\beta}_\mathbf{ri} \sim n\left(\boldsymbol{\beta}_\mathbf{r}, \boldsymbol{\Sigma}\right) \end{aligned} \tag{1}$$

where $\boldsymbol{\beta}_\mathbf{f}$ is a vector of "fixed coefficients" shared by all individuals (as would be the case in the generic linear regression model), $\boldsymbol{\beta}_\mathbf{ri}$ is an individual-specific coefficient vector (also called "random coefficients"), and the regression error $\boldsymbol{\varepsilon}_\mathbf{i}$ has the usual properties of the basic regression model. The new element in this model is the *hyper-distribution* for the random vector $\boldsymbol{\beta}_\mathbf{ri}$. Specifically, each individual "draws" her $\boldsymbol{\beta}_\mathbf{ri}$ from another normal density with "grand mean" $\boldsymbol{\beta}_\mathbf{r}$ and VCOV $\boldsymbol{\Sigma}$. The random coefficients are often interpreted as allowing for "heterogeneity in tastes" over individuals. A classical example would be the effect of characteristics of consumer goods (such as cars) on potential buyers. You might care a lot about a great sound system, but I may not care as much. Such, if your regression measures the maximum price a person would pay for a car, and "sound system quality" is one of the

regressors, it would make sense to ex ante allow for different marginal effects of this regressor over individuals.

Theoretically, one could aim at estimating all *N* vectors $\boldsymbol{\beta}_{\mathbf{ri}}$, but that would require *N* additional degrees of freedom, which may be infeasible in many applications. A more compact approach is to draw $\boldsymbol{\beta}_{\mathbf{ri}}$ from a shared hyper-distribution as shown in (1). That is the essence of hierarchical modeling.

It is often convenient to re-write (1) in terms of deviations of random coefficients from the grand mean, i.e.

$$\begin{aligned} &\mathbf{y_i} = \mathbf{X_{fi}}\boldsymbol{\beta}_{\mathbf{f}} + \mathbf{X_{ri}}\boldsymbol{\beta}_{\mathbf{r}} + \mathbf{X_{ri}}\boldsymbol{\alpha}_{\mathbf{ri}} + \boldsymbol{\varepsilon}_{\mathbf{i}} \qquad \text{with} \quad \boldsymbol{\alpha}_{\mathbf{ri}} = \boldsymbol{\beta}_{\mathbf{ri}} - \boldsymbol{\beta}_{\mathbf{r}} \\ &\boldsymbol{\alpha}_{\mathbf{ri}} \sim n(\mathbf{0}, \boldsymbol{\Sigma}) \end{aligned} \tag{2}$$

This makes it easier to recognize the modeling distribution of $\mathbf{y_i}$, *unconditional* on $\boldsymbol{\beta}_{\mathbf{ri}}$ as

$$\mathbf{y_i} \sim n(\boldsymbol{\mu}_{\mathbf{i}}, \mathbf{V_i}), \quad \text{with} \quad \boldsymbol{\mu}_{\mathbf{i}} = \mathbf{X_{fi}}\boldsymbol{\beta}_{\mathbf{f}} + \mathbf{X_{ri}}\boldsymbol{\beta}_{\mathbf{r}}, \qquad \mathbf{V_i} = \mathbf{X_{ri}}\boldsymbol{\Sigma}\mathbf{X}'_{\mathbf{ri}} + \sigma^2\mathbf{I_J}, \tag{3}$$

which will be convenient later in our estimation. Note that we know the exact form of this unconditional density ONLY because we specified normal distributions for both our main regression model and the hierarchical elements. For most other combination of base density plus hierarchical density (or densities), we would have to work with explicit integration.

The full model for the entire sample of *NxJ* observations can be written as

$$\begin{aligned} &\mathbf{y} = \mathbf{X_f}\boldsymbol{\beta}_{\mathbf{f}} + \mathbf{X_r}\tilde{\boldsymbol{\beta}}_{\mathbf{r}} + \boldsymbol{\varepsilon} \qquad \text{where} \\ &\underset{(NxJ)\,xk_f}{\mathbf{X_f}} = \begin{bmatrix} \mathbf{X_{f1}} \\ \mathbf{X_{f2}} \\ \vdots \\ \mathbf{X_{fN}} \end{bmatrix} \qquad \underset{(NxJ)\,x\,(Nxk_r)}{\mathbf{X_r}} = \begin{bmatrix} \mathbf{X_{r1}} & 0 & \cdots & 0 \\ 0 & \mathbf{X_{r2}} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \mathbf{X_{rN}} \end{bmatrix} \qquad \underset{(Nxk_r)\,x1}{\tilde{\boldsymbol{\beta}}_{\mathbf{r}}} = \begin{bmatrix} \boldsymbol{\beta}_{\mathbf{r1}} \\ \boldsymbol{\beta}_{\mathbf{r2}} \\ \vdots \\ \boldsymbol{\beta}_{\mathbf{rN}} \end{bmatrix} \qquad \boldsymbol{\varepsilon} = \begin{pmatrix} \boldsymbol{\varepsilon}_{\mathbf{1}} \\ \boldsymbol{\varepsilon}_{\mathbf{2}} \\ \vdots \\ \boldsymbol{\varepsilon}_{\mathbf{N}} \end{pmatrix} \sim n\left(0, \sigma^2\mathbf{I_{NxJ}}\right) \end{aligned} \tag{4}$$

Our goal is to derive the posterior distribution for $\boldsymbol{\beta}_{\mathbf{f}}$, $\boldsymbol{\beta}_{\mathbf{r}}$, $\boldsymbol{\Sigma}$, and $\sigma^2$. We are NOT generally interested in realizations of the individual coefficient vectors $\boldsymbol{\beta}_{\mathbf{ri}}$, although they will help us draw the main parameters of interest in our Gibbs Sampler. I recommend always collecting these draws & saving them for later, especially if you're planning to use the Chib method to evaluate the marginal likelihood.

The likelihood function for this model, unconditional on the random effects, can be written as

$$p\left(\mathbf{y}\,|\,\mathbf{X},\boldsymbol{\beta}_\mathbf{f},\boldsymbol{\beta}_\mathbf{r},\boldsymbol{\Sigma},\sigma^2\right)=$$
$$\prod_{i=1}^{N}\left\{\int_{\boldsymbol{\beta}_{\mathbf{ri}}}(2\pi)^{-J/2}\,\sigma^{-J/2}\exp\left(-\tfrac{1}{2\sigma^2}\left(\mathbf{y_i}-\left(\mathbf{X_{fi}}\boldsymbol{\beta}_\mathbf{f}+\mathbf{X_{ri}}\boldsymbol{\beta}_{\mathbf{ri}}\right)\right)'\left(\mathbf{y_i}-\left(\mathbf{X_{fi}}\boldsymbol{\beta}_\mathbf{f}+\mathbf{X_{ri}}\boldsymbol{\beta}_{\mathbf{ri}}\right)\right)\right)p\left(\boldsymbol{\beta}_{\mathbf{ri}}\right)d\boldsymbol{\beta}_{\mathbf{ri}}\right\} \tag{5}$$
$$\text{with}\qquad p\left(\boldsymbol{\beta}_{\mathbf{ri}}\right)=\left(2\pi\right)^{-k_r/2}\left|\boldsymbol{\Sigma}\right|^{-1/2}\exp\left(-\tfrac{1}{2}\left(\boldsymbol{\beta}_{\mathbf{ri}}-\boldsymbol{\beta}_\mathbf{r}\right)'\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\beta}_{\mathbf{ri}}-\boldsymbol{\beta}_\mathbf{r}\right)\right).$$

In HMs, the *LHF* will usually include an integral over the random elements. This form would make it difficult to break the posterior into "well-behaved" conditionals for all parameters of interest for most models. In our case, we have a base normal density plus a hierarchical normal, which allows us to write the *LHF* without an explicit integration (or, in other words, solve the integral in (5) analytically). Using our result from (3), we can write

$$p\left(\mathbf{y}\,|\,\mathbf{X},\boldsymbol{\beta}_\mathbf{f},\boldsymbol{\beta}_\mathbf{r},\boldsymbol{\Sigma},\sigma^2\right)=\prod_{i=1}^{N}\left(2\pi\right)^{-J/2}\left|\mathbf{V_i}\right|^{-1/2}\exp\left(-\tfrac{1}{2}\left(\mathbf{y_i}-\boldsymbol{\mu_i}\right)'\mathbf{V_i^{-1}}\left(\mathbf{y_i}-\boldsymbol{\mu_i}\right)\right)=$$
$$\left(2\pi\right)^{-NJ/2}\left|\mathbf{V_i}\right|^{-N/2}\exp\left(-\tfrac{1}{2}\sum_{i=1}^{N}\left(\mathbf{y_i}-\boldsymbol{\mu_i}\right)'\mathbf{V_i^{-1}}\left(\mathbf{y_i}-\boldsymbol{\mu_i}\right)\right) \tag{6}$$
$$\text{with}\ \ \mathbf{V_i}=\mathbf{X_{ri}}\boldsymbol{\Sigma}\mathbf{X}'_{\mathbf{ri}}+\sigma^2\mathbf{I_J},\qquad \boldsymbol{\mu_i}=\mathbf{X_{fi}}\boldsymbol{\beta}_\mathbf{f}+\mathbf{X_{ri}}\boldsymbol{\beta}_\mathbf{r}.$$

*The priors*

We use our usual combination of normal, inverse gamma, and *IW* for prior densities. Specifically:

$$\begin{aligned}
p\left(\boldsymbol{\beta}_\mathbf{f}\right)&=n\left(\boldsymbol{\mu}_{\mathbf{0f}},\mathbf{V}_{\mathbf{0f}}\right)\\
p\left(\boldsymbol{\beta}_\mathbf{r}\right)&=n\left(\boldsymbol{\mu}_{\mathbf{0r}},\mathbf{V}_{\mathbf{0r}}\right)\\
p\left(\boldsymbol{\Sigma}\right)&=IW\left(\nu_0,S_0\right)\\
p\left(\sigma^2\right)&=ig\left(\eta_0,\tau_0\right)
\end{aligned} \tag{7}$$

Note that I have switched the notation for the shape parameter in the inverse gamma from $\nu_0$ to $\eta_0$ to distinguish it from the degrees-of-freedom parameter in the Inverse Wishart. You have already encountered the explicit forms of these densities in previous sections.

*The posterior and Data Augmentation*

Generically, we can write the posterior kernel as

$$p\left(\boldsymbol{\beta}_\mathbf{f},\boldsymbol{\beta}_\mathbf{r},\boldsymbol{\Sigma},\sigma^2\,|\,\mathbf{y},\mathbf{X}\right)\propto p\left(\boldsymbol{\beta}_\mathbf{f}\right)p\left(\boldsymbol{\beta}_\mathbf{r}\right)p\left(\boldsymbol{\Sigma}\right)p\left(\sigma^2\right)p\left(\mathbf{y}\,|\,\boldsymbol{\beta}_\mathbf{f},\boldsymbol{\beta}_\mathbf{r},\boldsymbol{\Sigma},\sigma^2,\mathbf{X}\right) \tag{8}$$

For some of our parameters, this form would make it very awkward to derive "well-behaved" conditionals. Specifically, for our application it would be convenient to work with the individual random vectors $\boldsymbol{\beta}_{\mathbf{ri}}$. To accomplish this without running into the integral in (5) we'll employ "*data augmentation*" as discussed in module 5.

For this application we augment our joint posterior with the *n* random coefficient vectors $\boldsymbol{\beta}_{\mathbf{ri}}$, i.e.

$$\begin{aligned} &p\left(\boldsymbol{\beta}_{\mathbf{f}},\boldsymbol{\beta}_{\mathbf{r}},\boldsymbol{\beta}_{\mathbf{ri}}\left(i=1\cdots N\right),\boldsymbol{\Sigma},\sigma^2 \mid \mathbf{y},\mathbf{X}\right)\propto \\ &p\left(\boldsymbol{\beta}_{\mathbf{f}}\right)p\left(\boldsymbol{\beta}_{\mathbf{r}},\boldsymbol{\beta}_{\mathbf{ri}}\left(i=1\cdots N\right),\boldsymbol{\Sigma}\right)p\left(\sigma^2\right)p\left(\mathbf{y}\mid\boldsymbol{\beta}_{\mathbf{f}},\boldsymbol{\beta}_{\mathbf{r}},\boldsymbol{\beta}_{\mathbf{ri}}\left(i=1\cdots N\right),\boldsymbol{\Sigma},\sigma^2,\mathbf{X}\right)= \\ &p\left(\boldsymbol{\beta}_{\mathbf{f}}\right)\left(\prod_{i=1}^{n}p\left(\boldsymbol{\beta}_{\mathbf{ri}}\mid\boldsymbol{\beta}_{\mathbf{r}},\boldsymbol{\Sigma}\right)\right)p\left(\boldsymbol{\beta}_{\mathbf{r}}\right)p\left(\boldsymbol{\Sigma}\right)p\left(\sigma^2\right)p\left(\mathbf{y}\mid\boldsymbol{\beta}_{\mathbf{f}},\boldsymbol{\beta}_{\mathbf{r}},\boldsymbol{\beta}_{\mathbf{ri}}\left(i=1\cdots N\right),\boldsymbol{\Sigma},\sigma^2,\mathbf{X}\right) \end{aligned} \tag{9}$$

The goal is then to integrate out the effect of the augmented data, i.e.

$$p\left(\boldsymbol{\theta}\mid\mathbf{y}\right)=\int_{\boldsymbol{\beta}_{\mathbf{ri}},i=1\ldots N}p\left(\boldsymbol{\theta}\mid\mathbf{y},\boldsymbol{\beta}_{\mathbf{ri}}\right)p\left(\boldsymbol{\beta}_{\mathbf{ri}}\mid\mathbf{y}\right)d\boldsymbol{\beta}_{\mathbf{ri}} \tag{10}$$

where $\boldsymbol{\theta}$ includes all actual parameters. This can be accomplished in straightforward manner by making draws of $\boldsymbol{\beta}_{\mathbf{ri}}$, conditional on actual data and all other parameters, a part of the Gibbs Sampler.

It is customary to think of the last part on the right hand side of (8) as the *conditional LHF*, which can be written as

$$\begin{aligned} &p\left(\mathbf{y}\mid\mathbf{X},\boldsymbol{\beta}_{\mathbf{f}},\boldsymbol{\beta}_{\mathbf{ri}}\left(i=1\cdots n\right),\sigma^2\right)= \\ &\left(2\pi\right)^{-NJ/2}\sigma^{-NJ/2}\exp\left(-\tfrac{1}{2\sigma^2}\sum_{i=1}^{N}\left(\mathbf{y_i}-\left(\mathbf{X_{fi}}\boldsymbol{\beta}_{\mathbf{f}}+\mathbf{X_{ri}}\boldsymbol{\beta}_{\mathbf{ri}}\right)\right)'\left(\mathbf{y_i}-\left(\mathbf{X_{fi}}\boldsymbol{\beta}_{\mathbf{f}}+\mathbf{X_{ri}}\boldsymbol{\beta}_{\mathbf{ri}}\right)\right)\right) \end{aligned} \tag{11}$$

We will drop the $\left(i=1\cdots n\right)$ part henceforth for notational convenience. It will be clear from the context if we condition on all *n* vectors of $\boldsymbol{\beta}_{\mathbf{ri}}$ or a specific $\boldsymbol{\beta}_{\mathbf{ri}}$ associated with a single individual.

Also, with data augmentation, we work with an additional "prior", i.e. the hierarchical prior $\left(\prod_{i=1}^{N}p\left(\boldsymbol{\beta}_{\mathbf{ri}}\mid\boldsymbol{\beta}_{\mathbf{r}},\boldsymbol{\Sigma}\right)\right)$. This hierarchical prior is normally listed along with the other priors. The priors for the moments of the hierarchical density (i.e. the priors for $\boldsymbol{\beta}_{\mathbf{r}}$ and $\boldsymbol{\Sigma}$ in our case) are often called "hyper-priors".

The full posterior kernel for the *augmented model* can then be explicitly written as:

$$
\begin{aligned}
&p\left(\boldsymbol{\beta}_{\mathbf{f}},\boldsymbol{\beta}_{\mathbf{r}},\boldsymbol{\beta}_{\mathbf{ri}}\left(i=1\cdots N\right),\boldsymbol{\Sigma},\sigma^{2}\mid\mathbf{y},\mathbf{X}\right)\propto\\
&\exp\left(-\tfrac{1}{2}\left(\boldsymbol{\beta}_{\mathbf{f}}-\boldsymbol{\mu}_{\mathbf{f0}}\right)'\mathbf{V}_{\mathbf{f0}}^{-1}\left(\boldsymbol{\beta}_{\mathbf{f}}-\boldsymbol{\mu}_{\mathbf{f0}}\right)\right)*\exp\left(-\tfrac{1}{2}\left(\boldsymbol{\beta}_{\mathbf{r}}-\boldsymbol{\mu}_{\mathbf{r0}}\right)'\mathbf{V}_{\mathbf{r0}}^{-1}\left(\boldsymbol{\beta}_{\mathbf{r}}-\boldsymbol{\mu}_{\mathbf{r0}}\right)\right)*\\
&\left(\sigma^{2}\right)^{-\left(\nu_{0}+1\right)}\exp\left(-\frac{\tau_{0}}{\sigma^{2}}\right)*\left|\boldsymbol{\Sigma}\right|^{-\left(\nu_{0}+M+1\right)/2}\exp\left(-\frac{1}{2}tr\left(\mathbf{S}_{\mathbf{0}}\cdot\boldsymbol{\Sigma}^{-1}\right)\right)*\\
&\left|\boldsymbol{\Sigma}\right|^{-1/2}\exp\left(-\tfrac{1}{2}\sum_{i=1}^{N}\left(\boldsymbol{\beta}_{\mathbf{ri}}-\boldsymbol{\beta}_{\mathbf{r}}\right)'\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\beta}_{\mathbf{ri}}-\boldsymbol{\beta}_{\mathbf{r}}\right)\right)*\\
&\sigma^{-NJ/2}\exp\left(-\tfrac{1}{2\sigma^{2}}\sum_{i=1}^{N}\left(\mathbf{y}_{\mathbf{i}}-\left(\mathbf{X}_{\mathbf{fi}}\boldsymbol{\beta}_{\mathbf{f}}+\mathbf{X}_{\mathbf{ri}}\boldsymbol{\beta}_{\mathbf{ri}}\right)\right)'\left(\mathbf{y}_{\mathbf{i}}-\left(\mathbf{X}_{\mathbf{fi}}\boldsymbol{\beta}_{\mathbf{f}}+\mathbf{X}_{\mathbf{ri}}\boldsymbol{\beta}_{\mathbf{ri}}\right)\right)\right)
\end{aligned}
\tag{12}
$$

*The Gibbs Sampler*

We will first examine the GS for the fully augmented model. We start by drawing the vector of fixed coefficients $\boldsymbol{\beta}_{\mathbf{f}}$. The conditional posterior kernel is given as:

$$
\begin{aligned}
&p\left(\boldsymbol{\beta}_{\mathbf{f}}\mid\boldsymbol{\beta}_{\mathbf{ri}},\sigma^{2},\mathbf{y},\mathbf{X}\right)\propto\\
&\exp\left(-\tfrac{1}{2\sigma^{2}}\sum_{i=1}^{N}\left(\mathbf{y}_{\mathbf{i}}-\left(\mathbf{X}_{\mathbf{fi}}\boldsymbol{\beta}_{\mathbf{f}}+\mathbf{X}_{\mathbf{ri}}\boldsymbol{\beta}_{\mathbf{ri}}\right)\right)'\left(\mathbf{y}_{\mathbf{i}}-\left(\mathbf{X}_{\mathbf{fi}}\boldsymbol{\beta}_{\mathbf{f}}+\mathbf{X}_{\mathbf{ri}}\boldsymbol{\beta}_{\mathbf{ri}}\right)\right)+\left(\boldsymbol{\beta}_{\mathbf{f}}-\boldsymbol{\mu}_{\mathbf{f0}}\right)'\mathbf{V}_{\mathbf{f0}}^{-1}\left(\boldsymbol{\beta}_{\mathbf{f}}-\boldsymbol{\mu}_{\mathbf{f0}}\right)\right)
\end{aligned}
\tag{13}
$$

Letting $\tilde{\mathbf{y}}_{\mathbf{i}}=\mathbf{y}_{\mathbf{i}}-\mathbf{X}'_{\mathbf{ri}}\boldsymbol{\beta}_{\mathbf{ri}}$ we can write this as

$$
\begin{aligned}
&p\left(\boldsymbol{\beta}_{\mathbf{f}}\mid\boldsymbol{\beta}_{\mathbf{ri}},\sigma^{2},\mathbf{y},\mathbf{X}\right)\propto\\
&\exp\left(-\tfrac{1}{2\sigma^{2}}\sum_{i=1}^{N}\left(\tilde{\mathbf{y}}_{\mathbf{i}}-\mathbf{X}_{\mathbf{fi}}\boldsymbol{\beta}_{\mathbf{f}}\right)'\left(\tilde{\mathbf{y}}_{\mathbf{i}}-\mathbf{X}_{\mathbf{fi}}\boldsymbol{\beta}_{\mathbf{f}}\right)+\left(\boldsymbol{\beta}_{\mathbf{f}}-\boldsymbol{\mu}_{\mathbf{f0}}\right)'\mathbf{V}_{\mathbf{f0}}^{-1}\left(\boldsymbol{\beta}_{\mathbf{f}}-\boldsymbol{\mu}_{\mathbf{f0}}\right)\right)
\end{aligned}
\tag{14}
$$

This brings us back to the exact same form of this conditional posterior kernel for the basic regression model with independent priors. We can use the results presented there to derive the conditional posterior as:

$$
\boldsymbol{\beta}_{\mathbf{f}}\mid\boldsymbol{\beta}_{\mathbf{ri}},\sigma^{2},\mathbf{y},\mathbf{X}\sim n\left(\boldsymbol{\mu}_{\mathbf{1}},\mathbf{V}_{\mathbf{1}}\right)\qquad\text{with}\qquad\mathbf{V}_{\mathbf{1}}=\left(\mathbf{V}_{\mathbf{0}}^{-1}+\tfrac{1}{\sigma^{2}}\mathbf{X}'_{\mathbf{f}}\mathbf{X}_{\mathbf{f}}\right)^{-1}\text{ and }\quad\boldsymbol{\mu}_{\mathbf{1}}=\mathbf{V}_{\mathbf{1}}\left(\mathbf{V}_{\mathbf{0}}^{-1}\boldsymbol{\mu}_{\mathbf{0}}+\tfrac{1}{\sigma^{2}}\mathbf{X}'_{\mathbf{f}}\tilde{\mathbf{y}}\right)\tag{15}
$$

Next, it is convenient to draw the individual random vectors $\boldsymbol{\beta}_{\mathbf{ir}}$. This requires working with our data-augmented framework. We aim to draw $\boldsymbol{\beta}_{\mathbf{ir}}$ from $p\left(\boldsymbol{\beta}_{\mathbf{ri}}\mid\boldsymbol{\beta}_{\mathbf{f}},\boldsymbol{\beta}_{\mathbf{r}},\boldsymbol{\Sigma},\sigma^{2},\mathbf{y}_{\mathbf{i}},\mathbf{X}_{\mathbf{i}}\right)$. Note that only data corresponding to individual "*i*" are relevant. The likelihood component for individual *i*, conditional on $\boldsymbol{\beta}_{\mathbf{ri}}$, is given by

$$p\left(\mathbf{y_i} \mid \mathbf{X_i}, \boldsymbol{\beta_f}, \boldsymbol{\beta_{ri}}, \sigma^2\right) = (2\pi)^{-J/2}\, \sigma^{-J/2} \exp\left( -\tfrac{1}{2\sigma^2}\left(\mathbf{y_i} - \left(\mathbf{X_{fi}}\boldsymbol{\beta_f} + \mathbf{X_{ri}}\boldsymbol{\beta_{ri}}\right)\right)' \left(\mathbf{y_i} - \left(\mathbf{X_{fi}}\boldsymbol{\beta_f} + \mathbf{X_{ri}}\boldsymbol{\beta_{ri}}\right)\right)\right) \tag{16}$$

which is basically a "mini" regression model with *J* observations. The "prior" for $\boldsymbol{\beta_{ir}}$ is simply its hierarchical distribution, i.e. the last equation in (5). In other words, we need the $i^{th}$ component of the conditional *LHF* in (11) plus the hierarchical prior of $\boldsymbol{\beta_{ir}}$ for this step.

Using a similar "trick" as before and defining $\tilde{\mathbf{y}}_\mathbf{i} = \mathbf{y_i} - \mathbf{X_{fi}}\boldsymbol{\beta_f}$, we can write the conditional kernel for drawing $\boldsymbol{\beta_{ir}}$ as

$$p\left(\boldsymbol{\beta_{ri}} \mid \boldsymbol{\beta_f}, \boldsymbol{\beta_r}, \boldsymbol{\Sigma}, \sigma^2, \mathbf{y_i}, \mathbf{X_i}\right) \propto \exp\left( -\tfrac{1}{2}\left( \tfrac{1}{\sigma^2}\left(\tilde{\mathbf{y}}_\mathbf{i} - \mathbf{X_{ri}}\boldsymbol{\beta_{ri}}\right)' \left(\tilde{\mathbf{y}}_\mathbf{i} - \mathbf{X_{ri}}\boldsymbol{\beta_{ri}}\right) + \left(\boldsymbol{\beta_{ri}} - \boldsymbol{\beta_r}\right)' \boldsymbol{\Sigma}^{-1} \left(\boldsymbol{\beta_{ri}} - \boldsymbol{\beta_r}\right)\right)\right) \tag{17}$$

This looks exactly like the kernel for the basic regression model, except with prior mean $\boldsymbol{\mu_0}$ replaced by hierarchical mean $\boldsymbol{\beta_r}$ and prior variance $\mathbf{V_0}$ replaced by the hierarchical variance $\boldsymbol{\Sigma}$. We can thus immediately write the conditional density as

$$\begin{aligned} &p\left(\boldsymbol{\beta_{ri}} \mid \boldsymbol{\beta_f}, \boldsymbol{\beta_r}, \boldsymbol{\Sigma}, \sigma^2, \mathbf{y_i}, \mathbf{X_i}\right) \sim n\left(\boldsymbol{\mu_{ri1}}, \mathbf{V_{ri1}}\right) \text{ with} \\ &\mathbf{V_{ri1}} = \left(\boldsymbol{\Sigma}^{-1} + \tfrac{1}{\sigma^2}\mathbf{X}'_{\mathbf{ri}}\mathbf{X_{ri}}\right)^{-1} \\ &\boldsymbol{\mu_{ri1}} = \mathbf{V_{ri1}}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta_r} + \tfrac{1}{\sigma^2}\mathbf{X}'_{\mathbf{ri}}\tilde{\mathbf{y}}_\mathbf{i}\right) = \mathbf{V_{ri1}}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta_r} + \tfrac{1}{\sigma^2}\mathbf{X}'_{\mathbf{ri}}\left(\mathbf{y_i} - \mathbf{X_{fi}}\boldsymbol{\beta_f}\right)\right) \end{aligned} \tag{18}$$

Keep in mind that you will have to repeat this *N* times, for each of the *N* $\boldsymbol{\beta_{ri}}$'s.

Conditional on knowing all $\boldsymbol{\beta_{ri}}$ 's, drawing the vector of random coefficient means, $\boldsymbol{\beta_r}$ becomes especially simple in data-augmented setting. Upon examination of the full posterior kernel we realize that $\boldsymbol{\beta_r}$ only appears in its own prior and the hyper-distribution for $\boldsymbol{\beta_{ri}}$. Thus it's conditional posterior kernel is independent of the likelihood and data. We have:

$$\begin{aligned} &p\left(\boldsymbol{\beta_r} \mid \boldsymbol{\beta_{ri}}, \boldsymbol{\Sigma}\right) \propto \\ &\exp\left( -\tfrac{1}{2}\left( \left(\boldsymbol{\beta_r} - \boldsymbol{\mu_{r0}}\right)' \mathbf{V}_{\mathbf{r0}}^{-1} \left(\boldsymbol{\beta_r} - \boldsymbol{\mu_{r0}}\right) + \sum_{i=1}^{N} \left(\boldsymbol{\beta_{ri}} - \boldsymbol{\beta_r}\right)' \boldsymbol{\Sigma}^{-1} \left(\boldsymbol{\beta_{ri}} - \boldsymbol{\beta_r}\right)\right)\right) \end{aligned} \tag{19}$$

Using transformations analogous to the SUR model we can derive the conditional posterior for $\boldsymbol{\beta_r}$ as

$$\begin{aligned} &\mathrm{p}\left(\boldsymbol{\beta_r} \mid \boldsymbol{\beta_{ri}}, \boldsymbol{\Sigma}\right) = mvn\left(\boldsymbol{\mu_1}, \mathbf{V_1}\right) \qquad \text{where} \\ &\mathbf{V_1} = \left(\mathbf{V}_{\mathbf{r0}}^{-1} + n \cdot \boldsymbol{\Sigma}^{-1}\right)^{-1} \qquad \boldsymbol{\mu_1} = \mathbf{V_1}\left(\mathbf{V}_{\mathbf{r0}}^{-1}\boldsymbol{\mu_{r0}} + \boldsymbol{\Sigma}^{-1}\sum_{i=1}^{n}\boldsymbol{\beta_{ri}}\right) \end{aligned} \tag{20}$$

Now on to the variance terms. As mentioned before, it will be convenient to draw $\sigma^2$ conditional on the $\boldsymbol{\beta}_{\mathbf{ri}}$ vectors. The relevant components of the full posterior kernel yield:

$$p\left(\sigma^2 \mid \boldsymbol{\beta}_{\mathbf{f}}, \boldsymbol{\beta}_{\mathbf{ri}}, \mathbf{y}, \mathbf{X}\right) \propto$$
$$\left(\sigma^2\right)^{-(\nu_0+1)} \exp\left(-\frac{\tau_0}{\sigma^2}\right) \sigma^{-NJ/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(\mathbf{y_i}-\left(\mathbf{X_{fi}}\boldsymbol{\beta}_{\mathbf{f}}+\mathbf{X_{ri}}\boldsymbol{\beta}_{\mathbf{ri}}\right)\right)'\left(\mathbf{y_i}-\left(\mathbf{X_{fi}}\boldsymbol{\beta}_{\mathbf{f}}+\mathbf{X_{ri}}\boldsymbol{\beta}_{\mathbf{ri}}\right)\right)\right) = \tag{21}$$
$$\left(\sigma^2\right)^{\frac{NJ-2\nu_0-2}{2}} \exp\left(-\frac{1}{2\sigma^2}\left(2\tau_0+\sum_{i=1}^{N}\left(\mathbf{y_i}-\left(\mathbf{X_{fi}}\boldsymbol{\beta}_{\mathbf{f}}+\mathbf{X_{ri}}\boldsymbol{\beta}_{\mathbf{ri}}\right)\right)'\left(\mathbf{y_i}-\left(\mathbf{X_{fi}}\boldsymbol{\beta}_{\mathbf{f}}+\mathbf{X_{ri}}\boldsymbol{\beta}_{\mathbf{ri}}\right)\right)\right)\right)$$

This lets us immediately use the results from the basic regression model with independent priors, i.e.

$$\sigma^2 \mid \boldsymbol{\beta}_{\mathbf{f}}, \boldsymbol{\beta}_{\mathbf{ri}}, \mathbf{y}, \mathbf{X} \sim ig\left(\eta_1, \tau_1\right) \qquad \text{with}$$
$$\eta_1 = \frac{2\eta_0 + n}{2} \qquad \text{and} \qquad \tau_1 = \frac{2\eta_0 + \sum_{i=1}^{n}\left(\mathbf{y_i}-\left(\mathbf{X_{fi}}\boldsymbol{\beta}_{\mathbf{f}}+\mathbf{X_{ri}}\boldsymbol{\beta}_{\mathbf{ri}}\right)\right)'\left(\mathbf{y_i}-\left(\mathbf{X_{fi}}\boldsymbol{\beta}_{\mathbf{f}}+\mathbf{X_{ri}}\boldsymbol{\beta}_{\mathbf{ri}}\right)\right)}{2} \tag{22}$$

The final step in our GS is the draw of the hierarchical variance $\boldsymbol{\Sigma}$. We use again our data-augmented framework and our analogy to the SUR model. Using similar arguments as for the draw of the hierarchical mean $\boldsymbol{\beta}_{\mathbf{r}}$, we get

$$p\left(\boldsymbol{\Sigma} \mid \boldsymbol{\beta}_{\mathbf{ri}}, \boldsymbol{\beta}_{\mathbf{r}}\right) = IW\left(\nu_1, S_1\right) \qquad \text{where}$$
$$\nu_1 = \nu_0 + n \qquad S_1 = S_0 + \sum_{i=1}^{n}\left(\boldsymbol{\beta}_{\mathbf{ri}}-\boldsymbol{\beta}_{\mathbf{r}}\right)\left(\boldsymbol{\beta}_{\mathbf{ri}}-\boldsymbol{\beta}_{\mathbf{r}}\right)' \tag{23}$$

An implementation of this framework using artificial data, generated in script `mod6_data`, is given in Matlab script `mod6_HNRM_v1`. The GS is provided in function `gs_HNRM_v1`.

*An improved Gibbs Sampler*

While the Sampler described above is easy to implement, it generally leads to considerable efficiency losses (i.e. high autocorrelation in the GS, slow convergence – see Chib & Carlin, 1999, for details), especially for draws of $\boldsymbol{\beta}_{\mathbf{f}}$ and $\boldsymbol{\beta}_{\mathbf{r}}$ (this is evident from the results of `HNRM_v1`). Thus, for these two parameters we will stick with our non-augmented framework implicitly given in (8). This also implies that we will work with the unconditional *LHF* in (6).

Another way to look at this is from a paradigm of "efficient blocking", briefly discussed in Module 2. For the fully augmented model, the blocking is as follows:

1. Draw $\boldsymbol{\beta}_{\mathbf{f}}$ from $p\left(\boldsymbol{\beta}_{\mathbf{f}} \mid \boldsymbol{\beta}_{\mathbf{ri}}, \boldsymbol{\Sigma}, \sigma^2, \mathbf{y}, \mathbf{X}\right)$

2. Draw $\boldsymbol{\beta}_{\mathbf{ri}}$ from $p\left(\boldsymbol{\beta}_{\mathbf{ri}} \mid \boldsymbol{\beta}_{\mathbf{f}}, \boldsymbol{\beta}_{\mathbf{r}}, \boldsymbol{\Sigma}, \sigma^2, \mathbf{y}_{\mathbf{i}}, \mathbf{X}_{\mathbf{i}}\right) \quad \forall i$
3. Draw $\boldsymbol{\beta}_{\mathbf{r}}$ from $p\left(\boldsymbol{\beta}_{\mathbf{r}} \mid \{\boldsymbol{\beta}_{\mathbf{ri}}\}_{i=1}^{N}, \boldsymbol{\Sigma}\right)$
4. Draw $\sigma^2$ from $p\left(\sigma^2 \mid \boldsymbol{\beta}_{\mathbf{f}}, \boldsymbol{\beta}_{\mathbf{ri}}, \mathbf{y}, \mathbf{X}\right)$
5. Draw $\boldsymbol{\Sigma}$ from $p\left(\boldsymbol{\Sigma} \mid \boldsymbol{\beta}_{\mathbf{ri}}, \boldsymbol{\beta}_{\mathbf{r}}\right)$

Thus, even though $\boldsymbol{\beta}_{\mathbf{ri}}$ is the only posterior element that is drawn conditional on *all other* parameters, we have five separate blocks. This is because for the remaining blocks, all parameters involved are mutually conditional with respect to at least one other conditioning element. To be specific, $\boldsymbol{\beta}_{\mathbf{r}}$ is conditioned on $\boldsymbol{\Sigma}$, and vice versa, and $\boldsymbol{\beta}_{\mathbf{f}}$ is conditioned on $\sigma^2$ and vice versa.

We have learned that, in general, a more efficient GS results if we use as few blocks as possible. Here, we can exploit the fact that we know the form of the partially conditional densities $p\left(\boldsymbol{\beta}_{\mathbf{f}} \mid \boldsymbol{\beta}_{\mathbf{r}}, \boldsymbol{\Sigma}, \sigma^2, \mathbf{y}, \mathbf{X}\right)$, and $p\left(\boldsymbol{\beta}_{\mathbf{r}} \mid \boldsymbol{\beta}_{\mathbf{f}}, \sigma^2, \boldsymbol{\Sigma}, \mathbf{y}, \mathbf{X}\right)$. In other words, we can draw $\boldsymbol{\beta}_{\mathbf{f}}$, $\boldsymbol{\beta}_{\mathbf{r}}$ and $\boldsymbol{\beta}_{\mathbf{ri}}$ in two blocks as follows (see Chib and Carlin, 1999):

1. Draw $\boldsymbol{\beta}_{\mathbf{f}}$ and $\boldsymbol{\beta}_{\mathbf{ri}}$ from $p\left(\boldsymbol{\beta}_{\mathbf{f}}, \{\boldsymbol{\beta}_{\mathbf{ri}}\}_{i=1}^{N} \mid \boldsymbol{\beta}_{\mathbf{r}}, \sigma^2, \boldsymbol{\Sigma}, \mathbf{y}, \mathbf{X}\right)$ via:
    a. Draw $\boldsymbol{\beta}_{\mathbf{ri}}$ from $p\left(\{\boldsymbol{\beta}_{\mathbf{ri}}\}_{i=1}^{N} \mid \boldsymbol{\beta}_{\mathbf{f}}, \boldsymbol{\beta}_{\mathbf{r}}, \sigma^2, \boldsymbol{\Sigma}, \mathbf{y}_i, \mathbf{X}_i\right)$
    b. Draw $\boldsymbol{\beta}_{\mathbf{f}}$ from $p\left(\boldsymbol{\beta}_{\mathbf{f}} \mid \sigma^2, \boldsymbol{\beta}_{\mathbf{r}}, \boldsymbol{\Sigma}, \mathbf{y}, \mathbf{X}\right)$
2. Draw $\boldsymbol{\beta}_{\mathbf{r}}$ from $p\left(\boldsymbol{\beta}_{\mathbf{r}} \mid \boldsymbol{\beta}_{\mathbf{f}}, \sigma^2, \boldsymbol{\Sigma}, \mathbf{y}, \mathbf{X}\right)$

The variance terms are drawn as in the fully augmented model. This generally yields a more efficient posterior sampler.

The full posterior kernel for the *non*-augmented model can then be explicitly written as:

$$
\begin{aligned}
& p\left(\boldsymbol{\beta}_{\mathbf{f}}, \boldsymbol{\beta}_{\mathbf{r}}, \boldsymbol{\Sigma}, \sigma^2 \mid \mathbf{y}, \mathbf{X}\right) \propto \\
& \exp\left(-\tfrac{1}{2}\left(\boldsymbol{\beta}_{\mathbf{f}}-\boldsymbol{\mu}_{\mathbf{f0}}\right)' \mathbf{V}_{\mathbf{f0}}^{-1}\left(\boldsymbol{\beta}_{\mathbf{f}}-\boldsymbol{\mu}_{\mathbf{f0}}\right)\right) * \exp\left(-\tfrac{1}{2}\left(\boldsymbol{\beta}_{\mathbf{r}}-\boldsymbol{\mu}_{\mathbf{r0}}\right)' \mathbf{V}_{\mathbf{r0}}^{-1}\left(\boldsymbol{\beta}_{\mathbf{r}}-\boldsymbol{\mu}_{\mathbf{r0}}\right)\right) * \\
& \left(\sigma^2\right)^{-(\nu_0+1)} \exp\left(-\frac{\tau_0}{\sigma^2}\right) * |\boldsymbol{\Sigma}|^{-(\nu_0+M+1)/2} \exp\left(-\frac{1}{2} tr\left(\mathbf{S}_{\mathbf{0}} \cdot \boldsymbol{\Sigma}^{-1}\right)\right) * \\
& \left|\left(\mathbf{X}_{\mathbf{ri}} \boldsymbol{\Sigma} \mathbf{X}_{\mathbf{ri}}' + \sigma^2 \mathbf{I}_{\mathbf{J}}\right)\right|^{-N/2} \exp\left(-\tfrac{1}{2} \sum_{i=1}^{N}\left(\mathbf{y}_{\mathbf{i}}-\boldsymbol{\mu}_{\mathbf{i}}\right)' \left(\mathbf{X}_{\mathbf{ri}} \boldsymbol{\Sigma} \mathbf{X}_{\mathbf{ri}}' + \sigma^2 \mathbf{I}_{\mathbf{J}}\right)^{-1}\left(\mathbf{y}_{\mathbf{i}}-\boldsymbol{\mu}_{\mathbf{i}}\right)\right)
\end{aligned}
\tag{24}
$$

In contrast to the fully augmented model in (12) the hierarchical prior for $\boldsymbol{\beta}_{\mathbf{ri}}$ and the conditional likelihood, i.e. the last two lines in (12) have now been replaced by the unconditional likelihood. Note again that this option of switching between the data- augmented model and the original model for a sub-

set of parameters only exists because we happen to know the analytical form of the non-augmented likelihood for the normal-normal case. In other hierarchical settings that do not build on a normal likelihood function and / or do not use a normal hyper-distribution for random coefficients we would have to stick with the fully augmented model.

Letting again $\tilde{\mathbf{y}}_\mathbf{i} = \mathbf{y}_\mathbf{i} - \mathbf{X}'_{\mathbf{ri}}\boldsymbol{\beta}_\mathbf{r}$ we can write the conditional posterior kernel for $\boldsymbol{\beta}_\mathbf{f}$ as

$$p\left(\boldsymbol{\beta}_\mathbf{f} \mid \boldsymbol{\beta}_\mathbf{r}, \boldsymbol{\Sigma}, \sigma^2, \mathbf{y}, \mathbf{X}\right) \propto \exp\left( -\tfrac{1}{2}\left( \sum_{i=1}^{n} \left(\tilde{\mathbf{y}}_\mathbf{i} - \mathbf{X}_{\mathbf{fi}}\boldsymbol{\beta}_\mathbf{f}\right)' \mathbf{V}_\mathbf{i}^{-1} \left(\tilde{\mathbf{y}}_\mathbf{i} - \mathbf{X}_{\mathbf{fi}}\boldsymbol{\beta}_\mathbf{f}\right) + \left(\boldsymbol{\beta}_\mathbf{f} - \boldsymbol{\mu}_{\mathbf{f0}}\right)' \mathbf{V}_{\mathbf{f0}}^{-1} \left(\boldsymbol{\beta}_\mathbf{f} - \boldsymbol{\mu}_{\mathbf{f0}}\right) \right) \right) \tag{25}$$

with $\mathbf{V}_\mathbf{i} = \mathbf{X}_{\mathbf{ri}}\boldsymbol{\Sigma}\mathbf{X}'_{\mathbf{ri}} + \sigma^2\mathbf{I}_\mathbf{J}$

Using similar transformations as for the SUR model, we can derive the conditional kernel of $\boldsymbol{\beta}_\mathbf{f}$ as

$$\boldsymbol{\beta}_\mathbf{f} \mid \boldsymbol{\beta}_\mathbf{r}, \boldsymbol{\Sigma}, \sigma^2, \mathbf{y}, \mathbf{X} \sim n\left(\boldsymbol{\mu}_{\mathbf{f1}}, \mathbf{V}_{\mathbf{f1}}\right) \quad \text{with}$$

$$\mathbf{V}_{\mathbf{f1}} = \left( \mathbf{V}_{\mathbf{f0}}^{-1} + \sum_{i=1}^{n} \mathbf{X}'_{\mathbf{fi}} \mathbf{V}_\mathbf{i}^{-1} \mathbf{X}_{\mathbf{fi}} \right)^{-1} = \left( \mathbf{V}_{\mathbf{f0}}^{-1} + \sum_{i=1}^{n} \mathbf{X}'_{\mathbf{fi}} \left( \mathbf{X}_{\mathbf{ri}}\boldsymbol{\Sigma}\mathbf{X}'_{\mathbf{ri}} + \sigma^2\mathbf{I}_\mathbf{J} \right)^{-1} \mathbf{X}_{\mathbf{fi}} \right)^{-1} \tag{26}$$

$$\boldsymbol{\mu}_{\mathbf{f1}} = \mathbf{V}_{\mathbf{f1}} \left( \mathbf{V}_{\mathbf{f0}}^{-1}\boldsymbol{\mu}_{\mathbf{f0}} + \sum_{i=1}^{n} \mathbf{X}'_{\mathbf{fi}} \mathbf{V}_\mathbf{i}^{-1} \tilde{\mathbf{y}}_\mathbf{i} \right) = \left( \mathbf{V}_{\mathbf{f0}}^{-1}\boldsymbol{\mu}_{\mathbf{f0}} + \sum_{i=1}^{n} \mathbf{X}'_{\mathbf{fi}} \left( \mathbf{X}_{\mathbf{ri}}\boldsymbol{\Sigma}\mathbf{X}'_{\mathbf{ri}} + \sigma^2\mathbf{I}_\mathbf{J} \right)^{-1} \left( \mathbf{y}_\mathbf{i} - \mathbf{X}_{\mathbf{ri}}\boldsymbol{\beta}_\mathbf{r} \right) \right)$$

Next, we draw the hierarchical mean vector $\boldsymbol{\beta}_\mathbf{r}$. As for $\boldsymbol{\beta}_\mathbf{f}$, we can gain efficiency by NOT conditioning on $\boldsymbol{\beta}_{\mathbf{ri}}$. Letting $\tilde{\mathbf{y}}_\mathbf{i} = \mathbf{y}_\mathbf{i} - \mathbf{X}'_{\mathbf{fi}}\boldsymbol{\beta}_\mathbf{f}$ we can write the conditional posterior kernel as

$$p\left(\boldsymbol{\beta}_\mathbf{r} \mid \boldsymbol{\beta}_\mathbf{f}, \boldsymbol{\Sigma}, \sigma^2, \mathbf{y}, \mathbf{X}\right) \propto \exp\left( -\tfrac{1}{2}\left( \sum_{i=1}^{n} \left(\tilde{\mathbf{y}}_\mathbf{i} - \mathbf{X}_{\mathbf{ri}}\boldsymbol{\beta}_\mathbf{r}\right)' \mathbf{V}_\mathbf{i}^{-1} \left(\tilde{\mathbf{y}}_\mathbf{i} - \mathbf{X}_{\mathbf{ri}}\boldsymbol{\beta}_\mathbf{r}\right) + \left(\boldsymbol{\beta}_\mathbf{r} - \boldsymbol{\mu}_{\mathbf{r0}}\right)' \mathbf{V}_{\mathbf{r0}}^{-1} \left(\boldsymbol{\beta}_\mathbf{r} - \boldsymbol{\mu}_{\mathbf{r0}}\right) \right) \right) \tag{27}$$

Thus, in perfect analogy to $\boldsymbol{\beta}_\mathbf{f}$, we can write

$$\boldsymbol{\beta}_\mathbf{r} \mid \boldsymbol{\beta}_\mathbf{f}, \boldsymbol{\Sigma}, \sigma^2, \mathbf{y}, \mathbf{X} \sim n\left(\boldsymbol{\mu}_{\mathbf{r1}}, \mathbf{V}_{\mathbf{r1}}\right) \quad \text{with}$$

$$\mathbf{V}_{\mathbf{r1}} = \left( \mathbf{V}_{\mathbf{r0}}^{-1} + \sum_{i=1}^{n} \mathbf{X}'_{\mathbf{ri}} \mathbf{V}_\mathbf{i}^{-1} \mathbf{X}_{\mathbf{ri}} \right)^{-1} = \left( \mathbf{V}_{\mathbf{r0}}^{-1} + \sum_{i=1}^{n} \mathbf{X}'_{\mathbf{ri}} \left( \mathbf{X}_{\mathbf{ri}}\boldsymbol{\Sigma}\mathbf{X}'_{\mathbf{ri}} + \sigma^2\mathbf{I}_\mathbf{J} \right)^{-1} \mathbf{X}_{\mathbf{ri}} \right)^{-1} \tag{28}$$

$$\boldsymbol{\mu}_{\mathbf{r1}} = \mathbf{V}_{\mathbf{r1}} \left( \mathbf{V}_{\mathbf{r0}}^{-1}\boldsymbol{\mu}_{\mathbf{r0}} + \sum_{i=1}^{n} \mathbf{X}'_{\mathbf{ri}} \mathbf{V}_\mathbf{i}^{-1} \tilde{\mathbf{y}}_\mathbf{i} \right) = \left( \mathbf{V}_{\mathbf{r0}}^{-1}\boldsymbol{\mu}_{\mathbf{r0}} + \sum_{i=1}^{n} \mathbf{X}'_{\mathbf{ri}} \left( \mathbf{X}_{\mathbf{ri}}\boldsymbol{\Sigma}\mathbf{X}'_{\mathbf{ri}} + \sigma^2\mathbf{I}_\mathbf{J} \right)^{-1} \left( \mathbf{y}_\mathbf{i} - \mathbf{X}_{\mathbf{fi}}\boldsymbol{\beta}_\mathbf{f} \right) \right)$$

The next steps are conditional draws of $\boldsymbol{\beta}_{\mathbf{ri}}$, $\sigma^2$, and $\boldsymbol{\Sigma}$. For these we revert back to the fully augmented framework above.

An implementation of this semi-augmented framework using the same artificial data as above is given in Matlab script `mod6_HNRM_v2`. The GS is provided in function `gs_HNRM_v2`.

## HNRM application

Matlab scripts: `mod6_HNRM_outage`

Let's re-visit the commercial outage data, using a different regression model. Specifically, we draw a set of 50 size 2 and 3 firms, and regress their outage costs (in $1000) against the scenario attributes "weekday", "daytime", and outage length (in minutes). The data file is `outage_50firms`.

(We'll ignore for now that there is considerable censoring in our data since 28% of cost reports are zero.)

Since it is very likely that not all firms are affected equally by these attributes, we will specify them as random coefficients with hierarchical priors.

1. Run script `mod6_HNRM_outage` . Looking at the GS diagnostics it is obvious that the covariances in $\boldsymbol{\Sigma}$ are essentially zero and / or poorly identified. This suggests running a model without covariance terms for the VCOV of the random coefficients (but still with variances).

2. The No-covariance model is implemented in script `mod6_HNRM_outage_nocovs`. It is identical to the first script, except we now specify separate (but identical) priors for the diagonal elements of $\boldsymbol{\Sigma}$. We then use a slightly modified GS (called `gs_HNRM_nocovs`) that replaces draws of $\boldsymbol{\Sigma}$ from the *IW* with *kr* draws of variances from the *ig*. However, for efficiency purposes, we draw them in one block, i.e. we do not condition one variance on draws of the others.

   The diagnostics for this version are clearly improved.

A formal comparison of the two models (full $\boldsymbol{\Sigma}$ and diagonal $\boldsymbol{\Sigma}$) requires the computation or approximation of the marginal likelihood for each case. Here we will use Chib's (1995) method to simulate the marginal likelihood, adjusted for data augmented models as discussed in modules 4 and 5.

We first initiate the Chib routine for the unconstrained model (full $\boldsymbol{\Sigma}$) in script `mod6_HNRM_outage_chib`, which calls the function `gs_HNRM_chib`. The output from this script are the marginal likelihood (in log form), as well as the three separate components that feed into it, i.e. the prior evaluated at the posterior mean, the sample likelihood evaluated at the posterior mean, and the posterior itself evaluated at the posterior mean (all in log form). We then repeat this procedure for the constrained model via script `mod6_HNRM_outage_nocovs_chib` and function `gs_HNRM_nocovs_chib`.

As you can see from the log files the log-Bayes Factor for the unconstrained vs. the constrained model is approximately 7.3, given our prior settings. According to the Table for Bayes Factor interpretation provided in Kass and Raftery (1995), this assigns *strong support* to the full model, despite the negligible magnitudes for its covariance terms (a log-BF of 6 or higher indicates strong support, 12 or higher = decisive support). One thing that "hurts" the restricted model is its larger prior space for the variances –

we now need three separate ig priors compared to a single IW prior. As you can see from the Chib results, this imposes a considerable penalty on the no-covariance specification.

## Generating Posterior Predictive Densities with hierarchical models

**Option 1: Using the fully augmented model:**
When working with HMs, PPDs often require an additional simulation step to marginalize of the hierarchical elements. For our fully augmented HNRM, for example, the PPD for a hypothetical observation with attributes $\mathbf{x_p}$ takes the following form:

$$p(\hat{y}_p \mid \mathbf{y}) = \int_{\boldsymbol{\theta}} \left( \int_{\boldsymbol{\beta}_{\mathbf{ri}}} p(\hat{y}_p, \boldsymbol{\theta}, \boldsymbol{\beta}_{\mathbf{ri}} \mid \mathbf{y}) d\boldsymbol{\beta}_{\mathbf{ri}} \right) d\,\boldsymbol{\theta} = \int_{\boldsymbol{\theta}} \left( \int_{\boldsymbol{\beta}_{\mathbf{ri}}} p(\hat{y}_p \mid \boldsymbol{\theta}, \boldsymbol{\beta}_{\mathbf{ri}}, \mathbf{y}) p(\boldsymbol{\theta}, \boldsymbol{\beta}_{\mathbf{ri}} \mid \mathbf{y}) d\boldsymbol{\beta}_{\mathbf{ri}} \right) d\,\boldsymbol{\theta} =$$
$$\int_{\boldsymbol{\theta}} \left( \int_{\boldsymbol{\beta}_{\mathbf{ri}}} p(\hat{y}_p \mid \boldsymbol{\theta}, \boldsymbol{\beta}_{\mathbf{ri}}) p(\boldsymbol{\beta}_{\mathbf{ri}} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathbf{y}) d\boldsymbol{\beta}_{\mathbf{ri}} \right) d\,\boldsymbol{\theta} = \int_{\boldsymbol{\theta}} \left( \int_{\boldsymbol{\beta}_{\mathbf{ri}}} p(\hat{y}_p \mid \boldsymbol{\theta}, \boldsymbol{\beta}_{\mathbf{ri}}) p(\boldsymbol{\beta}_{\mathbf{ri}} \mid \boldsymbol{\theta}) d\boldsymbol{\beta}_{\mathbf{ri}} \right) p(\boldsymbol{\theta} \mid \mathbf{y}) d\,\boldsymbol{\theta} \quad (29)$$

where $\boldsymbol{\theta}$ captures all actual parameters. This suggests the following implementation:

1. For a given set of parameters $\boldsymbol{\theta}$ from the original GS, draw at least 1 $\boldsymbol{\beta}_{\mathbf{ri}}$ from its hierarchical prior $p(\boldsymbol{\beta}_{\mathbf{ri}} \mid \boldsymbol{\beta}_{\mathbf{r}}, \boldsymbol{\Sigma})$ (more draws will smoothen the PPD).
2. Draw a value of the predictive construct $\hat{y}_p$ from the conditional likelihood $p(\hat{y}_p \mid \boldsymbol{\theta}, \boldsymbol{\beta}_{\mathbf{ri}})$.
3. Repeat for all *R* draws of $\boldsymbol{\theta}$.

**Option 2: Using the partially augmented model:**
For our very special case of the normal hierarchical regression model, we can equivalently use:

$$p(\hat{y}_p \mid \mathbf{y}) = \int_{\boldsymbol{\theta}} p(\hat{y}_p, \boldsymbol{\theta} \mid \mathbf{y})\, d\,\boldsymbol{\theta} = \int_{\boldsymbol{\theta}} p(\hat{y}_p \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathbf{y}) d\,\boldsymbol{\theta} =$$
$$\int_{\boldsymbol{\theta}} p(\hat{y}_p \mid \boldsymbol{\beta}_{\mathbf{f}}, \boldsymbol{\beta}_{\mathbf{r}}, \boldsymbol{\Sigma}, \sigma^2) p(\boldsymbol{\beta}_{\mathbf{f}}, \boldsymbol{\beta}_{\mathbf{r}}, \boldsymbol{\Sigma}, \sigma^2 \mid \mathbf{y}) d\,\boldsymbol{\theta}$$

with the following implementation:

1. For a given set of parameters $\boldsymbol{\theta}$ from the original GS, draw one or more values of the predictive construct $\hat{y}_p$ from the un-augmented (original) likelihood $p(\hat{y}_p \mid \boldsymbol{\theta})$.
2. Repeat for all *R* draws of $\boldsymbol{\theta}$.

Script `mod6_HNRM_outage_predict` shows both versions.

## References:

Chib, Siddhartha and Bradley P. Carlin. 1999. "On MCMC sampling in hierarchical longitudinal models." *Statistics and Computing* 9, 17-26.

Kass, R. E. and A. E. Raftery. 1995. "Bayes Factors." *Journal of the American Statistical Association* 90, 773-795.