Multinomial Probit Models

KPT, Ch. 14

Matlab scripts: `mod12_MNP_I_data, mod12_MNP_I`
Matlab functions: `gs_MNP_I, GwG`

The MNP model applies to situations where individuals have to choose from >2 unordered choices, such as modes of transportation, brands of consumer products, or a menu of hypothetical policy scenarios in contingent experiments.

We will again use a latent data framework, but this time the latent construct is more firmly anchored in economic theory. Specifically, we will use a Random Utility Modeling (RUM) framework. In a RUM framework, we assume that person *i* derives utility from option *j*. This utility is modeled as the sum of an observable component and an error term. In the simplest case, the observed component is a linear function of choice attributes and coefficients, i.e.

$$U_{ij} = \mathbf{x}_{ij}'\boldsymbol{\beta} + \varepsilon_{ij} \qquad j = 1 \cdots J \qquad i = 1 \cdots N \tag{1}$$

In many applications $\mathbf{x}_{ij}$ will only include attributes corresponding to the j^{th} option. However, since not every individual may face the exact same set of options, we index $\mathbf{x}$ by *j* AND *i*. For example, in contingent experiments, different individuals will be given different choice menus BY DESIGN to enhance the properties of the resulting estimator.

It is also possible to introduce observed respondent characteristics in to the model via interactions with the choice attributes. This would be a second reason to index $\mathbf{x}$ by *j* and *i*. For simplicity, we will abstract from any socio-demographic interactions in this exposition. Another extension would allow the $\boldsymbol{\beta}$-vectors to vary over choices. However, this is only meaningful if every respondent faces the exact same set of options. We will abstract from this extension as well.

At the individual level we have

$$\mathbf{U}_i = \mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i \qquad \boldsymbol{\varepsilon}_i \sim n(\mathbf{0}, \boldsymbol{\Sigma}) \tag{2}$$

As discussed in Train (2003), Ch. 5., in absence of any researcher-imposed structural restrictions on $\boldsymbol{\Sigma}$ the model needs to be normalized for *level* (adding a constant to all utilities won't change the observed outcome) and *scale* (multiplying all utilities by the same scalar won't change the observed outcome). The first can be accomplished by declaring the utility for one of the alternatives (say the first) as "baseline", and differencing all other utilities with respect to that baseline. Naturally, this only makes sense if the exact same baseline alternative appears in all choice menus for all individuals. If this is not the case, the researcher needs to make arbitrary structural restrictions on $\boldsymbol{\Sigma}$ such as declaring it an identity matrix.

In choice experiments, the baseline is often the "status quo" alternative, for which the observed drivers of utility are often set to zero, i.e $\mathbf{x_{i1}} = 0 \ \forall i$. We can then express the model in terms of utility differences from the baseline, i.e.

$$U_{ij}^{*} = \mathbf{x_{ij}'\beta} + \varepsilon_{ij}^{*} \quad j = 2 \cdots J \qquad i = 1 \cdots n \qquad \varepsilon_{ij}^{*} = \varepsilon_{ij} - \varepsilon_{i1} \tag{3}$$

Note: If the baseline scenario has meaningful settings for $\mathbf{x_{ij}}$, we would compute the utility difference as

$$U_{ij}^{*} = \left(\mathbf{x_{ij}} - \mathbf{x_{i1}}\right)' \boldsymbol{\beta} + \varepsilon_{ij}^{*} \quad j = 2 \cdots J \qquad i = 1 \cdots n \qquad \varepsilon_{ij}^{*} = \varepsilon_{ij} - \varepsilon_{i1} \tag{4}$$

The system of *J-1* random utility differences for person *i* can be written as

$$\mathbf{U_i^{*}} = \mathbf{X_i}\boldsymbol{\beta} + \boldsymbol{\varepsilon_i^{*}} \tag{5}$$

Train (2003) shows a convenient "trick" to quickly compute the variance matrix for the differenced errors. Assume you start out with *J* alternatives (including the baseline). After differencing, you're left with *J-1* utility differences. Thus, declare a (*J-1*) by (*J-1*) identity matrix and "squeeze in" an extra column of "-1"'s in the position of the original baseline alternative. For example, if *J*=4 and the first alternative is the baseline, the resulting *Differencing Matrix* becomes

$$\mathbf{D} = \begin{bmatrix} -1 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ -1 & 0 & 0 & 1 \end{bmatrix} \tag{6}$$

and the differenced errors have the following distribution:

$$\boldsymbol{\varepsilon_i^{*}} = \mathbf{D}\boldsymbol{\varepsilon_i} \sim n\left(\mathbf{0}, \boldsymbol{\Sigma}^{*}\right) \qquad \boldsymbol{\Sigma}^{*} = \mathbf{D}\boldsymbol{\Sigma}\mathbf{D}' \tag{7}$$

For normalization with respect to *scale*, one of the variance elements in $\boldsymbol{\Sigma}^{*}$ needs to be fixed. The standard approach is to set $\sigma_{11} = 1$, as we did for the Probit and selection case. Thus, in theory, $\boldsymbol{\Sigma}^{*}$ can include $\left(\left(J-1\right)J/2\right)-1$ free elements. Train (2003) shows in detail how the elements in $\boldsymbol{\Sigma}^{*}$ relate to the elements in the original $\boldsymbol{\Sigma}$. He points out that while $\boldsymbol{\Sigma}^{*}$ still allows for substitution patterns (and thus overcomes the restrictive IIA requirement for logit-type models) , no intuition can be gained from inspection of its estimated elements with respect to the original variances and covariances. This subtle but important point is often missed in applied research.

Keep in mind that for applications with *changing choice options across respondents*, it is not meaningful to ex ante specify an unrestricted $\boldsymbol{\Sigma}$ - matrix, since the definition of "option 1, 2, 3, etc" is not consistent over respondents. In that case, it may be preferable to set $\boldsymbol{\Sigma} = \mathbf{I}$ as in the seminal paper by Hausman and

Wise (1978). This automatically normalizes the model for level and scale, and you do not need to ex ante work with utility differences.

Continuing, however, with the fully general case, the latent model for the entire sample emerges as

$$\mathbf{U}^* = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}^*$$

$$\mathbf{U}^* = \underset{(n^*(J-1))x1}{\begin{bmatrix} \mathbf{U}_1^* \\ \mathbf{U}_2^* \\ \vdots \\ \mathbf{U}_n^* \end{bmatrix}} \qquad \mathbf{X} = \underset{(n^*(J-1))xk}{\begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_n \end{bmatrix}} \qquad \boldsymbol{\varepsilon}^* = \begin{bmatrix} \boldsymbol{\varepsilon}_1^* \\ \boldsymbol{\varepsilon}_2^* \\ \vdots \\ \boldsymbol{\varepsilon}_n^* \end{bmatrix} \sim n(\mathbf{0}, \boldsymbol{\Omega}) \qquad \text{with} \tag{8}$$

$$\boldsymbol{\Omega} = \underset{(n^*(J-1))x(n^*(J-1))}{\begin{bmatrix} \boldsymbol{\Sigma}^* & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}^* & \cdots & \mathbf{0} \\ \mathbf{0} & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{\Sigma}^* \end{bmatrix}}$$

For each respondent we observe y_i, i.e. a scalar denoting the actual choice (or – equivalently – a *J* by 1 vector with zeros and a "1" for the chosen option). We can relate the observed choice index to latent utility differences as follows:

$$\begin{aligned} y_i = 1 \quad &\text{if} \qquad \max\left\{U_{ij}^*\right\}_{j=2}^{J} \le 0 \\ y_i = k \quad &\text{if} \qquad \max\left\{0, \left\{U_{ij}^*\right\}_{j=2}^{J}\right\} = U_{ik}^* \end{aligned} \tag{9}$$

An individual's contribution to the likelihood function can thus be expressed as a multivariate normal *cdf*, with choice-specific truncation bounds. For example, a choice of "1" implies:

$$\Pr(y_i = 1) = \Pr\begin{pmatrix} U_{i2}^* \le 0 \\ U_{i3}^* \le 0 \\ \vdots \\ U_{iJ}^* \le 0 \end{pmatrix} = \begin{pmatrix} \varepsilon_{i2}^* \le -\mathbf{x}_{\mathbf{i1}}'\boldsymbol{\beta} \\ \varepsilon_{i3}^* \le -\mathbf{x}_{\mathbf{i2}}'\boldsymbol{\beta} \\ \vdots \\ \varepsilon_{iJ}^* \le -\mathbf{x}_{\mathbf{iJ}}'\boldsymbol{\beta} \end{pmatrix} = \Phi\left(\mathbf{0}, \boldsymbol{\Sigma}^*; R_1\right) \tag{10}$$

where (with slight abuse of notation) in this case $\Phi(.)$ denotes the *cdf* of the truncated multivariate normal density with mean **0** and variance matrix $\boldsymbol{\Sigma}^*$, and truncation region R_j implicitly defined by the condition $(y_i = j)$. Generically, the likelihood contribution for the *ith* individual can be expressed as

$$p(y_i \mid \mathbf{X_i}, \boldsymbol{\beta}, \boldsymbol{\Sigma}) = \sum_{j=1}^{J}\left(\Phi\left(\mathbf{0}, \boldsymbol{\Sigma}^*; R_j\right) I(y_i = j)\right) \tag{11}$$

leading to the sample likelihood function

$$p(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\Sigma}) = \prod_{i=1}^{n} \left(\sum_{j=1}^{J} \left(\Phi\left(\mathbf{0}, \boldsymbol{\Sigma}^{*}; R_j\right) I\left(y_i = j\right) \right) \right) \tag{12}$$

The evaluation of the multivariate truncated normal terms is a major challenge in classical estimation. Usually, this is accomplished via simulation methods such as the GHK (after Geweke, Hajivassiliou, Keane – for a good discussion see Train (2003)). In Bayesian estimation, we can again resort to latent variable techniques to facilitate posterior simulation.

First the priors: We'll assign the usual normal prior to $\boldsymbol{\beta}$ and a constrained IW prior for $\boldsymbol{\Sigma}^{*}$ as we did for the selection model. The IW prior is adjusted for the *(J-1)* by *(J-1)* dimension of $\boldsymbol{\Sigma}^{*}$:

$$\begin{aligned}
&p\left(\boldsymbol{\beta}, \boldsymbol{\Sigma}^{*}\right) = p(\boldsymbol{\beta})\, p\left(\boldsymbol{\Sigma}^{*}\right) \quad \text{where} \\
&p(\boldsymbol{\beta}) = (2\pi)^{-k/2} \left|\mathbf{V_0}\right|^{-1/2} \exp\left(-\tfrac{1}{2} (\boldsymbol{\beta} - \boldsymbol{\mu_0})' \, \mathbf{V_0^{-1}} (\boldsymbol{\beta} - \boldsymbol{\mu_0}) \right) \\
&p\left(\boldsymbol{\Sigma}^{*}\right) = \left(2^{\nu_0 (J-1)/2} \pi^{(J-1)(J-2)/4} \prod_{i=1}^{J-1} \Gamma\left(\frac{\nu_0 + 1 - i}{2} \right) \right)^{-1} * \left|\mathbf{S_0}\right|^{\nu_0/2} \left|\boldsymbol{\Sigma}^{*}\right|^{-(\nu_0 + J)/2} \exp\left(-\frac{1}{2} tr\left(\mathbf{S_0} \cdot \left(\boldsymbol{\Sigma}^{*}\right)^{-1} \right) \right) I\left(\Sigma_{11}^{*} = 1\right)
\end{aligned} \tag{13}$$

Combining the priors with the likelihood, and dropping all terms that are multiplicatively unrelated to our parameters of interest yields the posterior kernel

$$\begin{aligned}
&p\left(\boldsymbol{\beta}, \boldsymbol{\Sigma}^{*} \mid \mathbf{y}, \mathbf{X}\right) \propto \\
&\left|\boldsymbol{\Sigma}^{*}\right|^{-(\nu_0 + J)/2} \exp\left(-\frac{1}{2} tr\left(\mathbf{S_0} \cdot \left(\boldsymbol{\Sigma}^{*}\right)^{-1} \right) \right) \exp\left(-\tfrac{1}{2} \left((\boldsymbol{\beta} - \boldsymbol{\mu_0})' \, \mathbf{V_0^{-1}} (\boldsymbol{\beta} - \boldsymbol{\mu_0}) \right) \right) * \\
&\prod_{i=1}^{n} \left(\sum_{j=1}^{J} \left(\Phi\left(\mathbf{0}, \boldsymbol{\Sigma}^{*}; R_j\right) I\left(y_i = j\right) \right) \right)
\end{aligned} \tag{14}$$

The augmented joint posterior takes the generic form of

$$p\left(\boldsymbol{\beta}, \boldsymbol{\Sigma}^{*}, \mathbf{U}^{*} \mid \mathbf{y}, \mathbf{X}\right) \propto p(\boldsymbol{\beta})\, p\left(\boldsymbol{\Sigma}^{*}\right) p\left(\mathbf{U}^{*} \mid \boldsymbol{\beta}, \boldsymbol{\Sigma}^{*}, \mathbf{X}\right) p\left(\mathbf{y} \mid \boldsymbol{\beta}, \boldsymbol{\Sigma}^{*}, \mathbf{X}, \mathbf{U}^{*}\right) \tag{15}$$

The first augmentation term, $p\left(\mathbf{U}^{*} \mid \boldsymbol{\beta}, \boldsymbol{\Sigma}^{*}, \mathbf{X}\right)$, is a simple product of individual-specific multivariate normal density terms similar to the SUR model, i.e.

$$p\left(\mathbf{U}^{*}\mid\boldsymbol{\beta},\boldsymbol{\Sigma}^{*},\mathbf{X}\right)=\prod_{i=1}^{n}\left(2\pi\right)^{-1}\left|\boldsymbol{\Sigma}^{*}\right|^{-1/2}\exp\left(-\tfrac{1}{2}\left(\mathbf{U}_{\mathbf{i}}^{*}-\mathbf{X}_{\mathbf{i}}\boldsymbol{\beta}\right)'\left(\boldsymbol{\Sigma}^{*}\right)^{-1}\left(\mathbf{U}_{\mathbf{i}}^{*}-\mathbf{X}_{\mathbf{i}}\boldsymbol{\beta}\right)\right)=$$
$$\left(2\pi\right)^{-n}\left|\boldsymbol{\Sigma}^{*}\right|^{-n/2}\exp\left(-\tfrac{1}{2}\sum_{i=1}^{n}\left(\mathbf{U}_{\mathbf{i}}^{*}-\mathbf{X}_{\mathbf{i}}\boldsymbol{\beta}\right)'\left(\boldsymbol{\Sigma}^{*}\right)^{-1}\left(\mathbf{U}_{\mathbf{i}}^{*}-\mathbf{X}_{\mathbf{i}}\boldsymbol{\beta}\right)\right) \tag{16}$$

The second term of the augmented data density can be written as

$$p\left(\mathbf{y}\mid\boldsymbol{\beta},\boldsymbol{\Sigma}^{*},\mathbf{X},\mathbf{U}^{*}\right)=p\left(\mathbf{y}\mid\mathbf{U}^{*}\right)=\prod_{i=1}^{n}\left(\begin{array}{l}I\left(y_i=1\right)I\left(\max\left\{U_{ij}^{*}\right\}_{j=2}^{J}\le 0\right)+\\ \sum_{k=2}^{J}I\left(y_i=k\right)I\left(\max\left\{0,\left\{U_{ij}^{*}\right\}_{j=2}^{J}\right\}=U_{ik}^{*}\right)\end{array}\right)$$

The augmented joint posterior can now be explicitly written as

$$p\left(\boldsymbol{\beta},\boldsymbol{\Sigma}^{*},\mathbf{U}^{*}\mid\mathbf{y},\mathbf{X}\right)\propto$$
$$\left|\boldsymbol{\Sigma}^{*}\right|^{-\left(\nu_0+J\right)/2}\exp\left(-\frac{1}{2}tr\left(\mathbf{S_0}\cdot\left(\boldsymbol{\Sigma}^{*}\right)^{-1}\right)\right)\exp\left(-\tfrac{1}{2}\left(\left(\boldsymbol{\beta}-\boldsymbol{\mu}_{\mathbf{0}}\right)'\mathbf{V_0^{-1}}\left(\boldsymbol{\beta}-\boldsymbol{\mu}_{\mathbf{0}}\right)\right)\right)*$$
$$\left(2\pi\right)^{-n}\left|\boldsymbol{\Sigma}^{*}\right|^{-n/2}\exp\left(-\tfrac{1}{2}\sum_{i=1}^{n}\left(\mathbf{U}_{\mathbf{i}}^{*}-\mathbf{X}_{\mathbf{i}}\boldsymbol{\beta}\right)'\left(\boldsymbol{\Sigma}^{*}\right)^{-1}\left(\mathbf{U}_{\mathbf{i}}^{*}-\mathbf{X}_{\mathbf{i}}\boldsymbol{\beta}\right)\right)*$$
$$\prod_{i=1}^{n}\left(I\left(y_i=1\right)I\left(\max\left\{U_{ij}^{*}\right\}_{j=2}^{J}\le 0\right)+\sum_{k=2}^{J}I\left(y_i=k\right)I\left(U_{ik}^{*}>\max\left\{0,U_{i-k}^{*}\right\}\right)\right)$$

The conditional posterior kernel for $\boldsymbol{\beta}$ now takes the following form:

$$p\left(\boldsymbol{\beta}\mid\boldsymbol{\Sigma}^{*},\mathbf{U}^{*},\mathbf{X}\right)\propto\exp\left(-\tfrac{1}{2}\left(\left(\boldsymbol{\beta}-\boldsymbol{\mu}_{\mathbf{0}}\right)'\mathbf{V_0^{-1}}\left(\boldsymbol{\beta}-\boldsymbol{\mu}_{\mathbf{0}}\right)+\sum_{i=1}^{n}\left(\mathbf{U}_{\mathbf{i}}^{*}-\mathbf{X}_{\mathbf{i}}\boldsymbol{\beta}\right)'\left(\boldsymbol{\Sigma}^{*}\right)^{-1}\left(\mathbf{U}_{\mathbf{i}}^{*}-\mathbf{X}_{\mathbf{i}}\boldsymbol{\beta}\right)\right)\right) \tag{17}$$

This is again equivalent to the conditional posterior for the basic SUR model, and we can immediately derive the conditional posterior moments as:

$$\boldsymbol{\beta}\mid\boldsymbol{\Sigma}^{*},\mathbf{U}^{*},\mathbf{X}\sim n\left(\boldsymbol{\mu}_{\mathbf{1}},\mathbf{V}_{\mathbf{1}}\right)\ \text{with}\quad\mathbf{V}_{\mathbf{1}}=\left(\mathbf{V_0^{-1}}+\sum_{i=1}^{n}\mathbf{X}_{\mathbf{i}}'\left(\boldsymbol{\Sigma}^{*}\right)^{-1}\mathbf{X}_{\mathbf{i}}\right)^{-1}\ \text{and}$$
$$\boldsymbol{\mu}_{\mathbf{1}}=\mathbf{V}_{\mathbf{1}}\left(\mathbf{V_0^{-1}}\boldsymbol{\mu}_{\mathbf{0}}+\sum_{i=1}^{n}\mathbf{X}_{\mathbf{i}}'\left(\boldsymbol{\Sigma}^{*}\right)^{-1}\mathbf{U}_{\mathbf{i}}^{*}\right) \tag{18}$$

For the conditional posterior of $\boldsymbol{\Sigma}^{*}$ we have

$$p\left(\boldsymbol{\Sigma}^{*} \mid \boldsymbol{\beta}, \mathbf{U}^{*}, \mathbf{X}\right) \propto$$
$$\left|\boldsymbol{\Sigma}^{*}\right|^{-(v_0+J+n)/2} \exp\left(-\tfrac{1}{2}\left(tr\left(\mathbf{S_0} \cdot \left(\boldsymbol{\Sigma}^{*}\right)^{-1}\right)+\sum_{i=1}^{n}\left(\mathbf{U_i^{*}}-\mathbf{X_i}\boldsymbol{\beta}\right)'\left(\boldsymbol{\Sigma}^{*}\right)^{-1}\left(\mathbf{U_i^{*}}-\mathbf{X_i}\boldsymbol{\beta}\right)\right)\right) I\left(\Sigma_{11}^{*}=0\right) \tag{19}$$

This is again analogous to the SUR model (except for the variance restriction), leading to

$$\boldsymbol{\Sigma}^{*} \mid \boldsymbol{\beta}, \mathbf{U}^{*}, \mathbf{X} \sim IW\left(v_1, \mathbf{S_1}\right) I\left(\Sigma_{11}^{*}=0\right) \qquad \text{with}$$
$$v_1 = v_0 + n \qquad \mathbf{S_1} = \mathbf{S_0} + \sum_{i=1}^{n}\left(\mathbf{U_i^{*}}-\mathbf{X_i}\boldsymbol{\beta}\right)\left(\mathbf{U_i^{*}}-\mathbf{X_i}\boldsymbol{\beta}\right)' \tag{20}$$

As for previously visited models, we can again use Nobile's (2000) technique (or equivalent methods) to draw the constrained $\boldsymbol{\Sigma}^{*}$ matrix.

The tricky part in the posterior simulation of the MNP model lies in the draws of the latent utility differences. For a given individual, the conditional posterior takes the form of

$$p\left(\mathbf{U_i^{*}} \mid \boldsymbol{\beta}, \boldsymbol{\Sigma}^{*}, y_i, \mathbf{X_i}\right) \propto$$
$$\exp\left(-\tfrac{1}{2}\left(\mathbf{U_i^{*}}-\mathbf{X_i}\boldsymbol{\beta}\right)'\left(\boldsymbol{\Sigma}^{*}\right)^{-1}\left(\mathbf{U_i^{*}}-\mathbf{X_i}\boldsymbol{\beta}\right)\right) * \left(I\left(\max\left\{U_{ij}^{*}\right\}_{j=2}^{J} \le 0\right)+\sum_{k=2}^{J} I\left(U_{ik}^{*} > \max\left\{0, \mathbf{U_{i\text{-}k}^{*}}\right\}\right)\right) \tag{21}$$

This essentially denotes a multivariate normal density with mean $\mathbf{X_i}\boldsymbol{\beta}$ and variance $\boldsymbol{\Sigma}^{*}$, truncated to a specific region defined by the relative magnitude of the utility terms. For example, if the first condition in (21) holds, i.e. $\max\left\{U_{ij}^{*}\right\}_{j=2}^{J} \le 0$, the truncation region is given by the following *J-2* jointly binding conditions:

$$R_1 : \begin{bmatrix} -\infty < U_{i2}^{*} \le 0 \\ -\infty < U_{i3}^{*} \le 0 \\ \vdots \\ -\infty < U_{iJ}^{*} \le 0 \end{bmatrix} \tag{22}$$

If option *k* yields the highest utility difference, the *J-1* dimensional truncation region becomes

$$R_k : \begin{bmatrix} -\infty < U_{i2}^* < U_{ik}^* \\ -\infty < U_{i3}^* < U_{ik}^* \\ \vdots \\ 0 < U_{ik}^* < \infty \\ \vdots \\ -\infty < U_{iJ}^* \le U_{ik}^* \end{bmatrix} \Rightarrow \begin{bmatrix} -\infty < U_{i2}^* - U_{ik}^* < 0 \\ -\infty < U_{i3}^* - U_{ik}^* < 0 \\ \vdots \\ -\infty < -U_{ik}^* < 0 \\ \vdots \\ -\infty < U_{iJ}^* - U_{ik}^* < 0 \end{bmatrix} \tag{23}$$

It is convenient to work again with differencing matrices. Let $\mathbf{U_i^*}$ be the *J-1* by 1 vector of original utility differences (after normalizing for level by subtracting the, say, first alternative's utility). Assume we observe $y_i = k$. We can then define differencing matrix $\mathbf{D_k}$ and a vector of second-differenced utilities $\mathbf{U_{ik}^*}$ s.t.

$$\mathbf{U_{ik}^*} = \begin{bmatrix} U_{i2}^* - U_{ik}^* \\ U_{i3}^* - U_{ik}^* \\ \vdots \\ -U_{ik}^* \\ \vdots \\ U_{iJ}^* - U_{ik}^* \end{bmatrix} = \mathbf{D_k U_i^*} \qquad \text{with} \qquad \mathbf{D_k} = \begin{bmatrix} 1 & 0 & \cdots & -1 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & -1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & -1 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & -1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & -1 & \cdots & 1 & 0 \\ 0 & 0 & \cdots & -1 & \cdots & 0 & 1 \end{bmatrix} \tag{24}$$

In this case $\mathbf{D_k}$ is a *J-1* by *J-1* identity matrix with a column of "-1's" *replacing* the column corresponding to the chosen alternative. Each choice of a different alternative implies a different $\mathbf{D_k}$ matrix. If $y_i = 1$, $\mathbf{D_k}$ is simply the identity matrix, as is evident from (22).

We can then relate the (untruncated) density of $\mathbf{U_i^*}$ to the (untruncated) density $\mathbf{U_{ik}^*}$ as follows:

$$\begin{aligned} &\mathbf{U_i^*} \sim n\left(\mathbf{X_i\beta}, \mathbf{\Sigma^*}\right) \qquad \rightarrow \qquad \mathbf{U_{ik}^*} = \mathbf{D_k U_i^*} \sim n\left(\mathbf{\mu_{ik}}, \mathbf{\Sigma_k^*}\right) \qquad \text{with} \\ &\mathbf{\mu_{ik}} = \mathbf{D_k X_i \beta}, \quad \mathbf{\Sigma_k^*} = \mathbf{D_k \Sigma^* D_k'} \end{aligned} \tag{25}$$

And the objective becomes to draw from

$$\mathbf{U_{ik}^*} \mid \mathbf{\beta}, \mathbf{\Sigma^*}, y_i, \mathbf{X_i} \sim n\left(\mathbf{\mu_{ik}}, \mathbf{\Sigma_k^*}, R_k\right) \tag{26}$$

where R_k signifies the truncation region given in (23). As discussed e.g. in Geweke (1991) , Koop (2003), and Koop et al. (2007), this can be accomplishes by breaking the (untruncated) density in (26) into univariate conditional densities, and then drawing the elements of $\mathbf{U_{ik}^*}$ one by one subject to their

respective truncation bounds. (Use `tnormrnd_robert_single` for one-sided truncation, and `tnormrnd` for double-sided truncation for best speed).

Since we don't observe an actual element from $\mathbf{U}_{\mathbf{ik}}^{*} \mid \boldsymbol{\beta}, \boldsymbol{\Sigma}^{*}, y_i, \mathbf{X}_{\mathbf{i}}; R_k$ we need to run a small "Gibbs-within-Gibbs" with, say, 10-30 iterations, to obtain the sequence of conditional draws. We treat all but the last set of draws as "burn-ins". A starting value of "0" can be used for the first element to get the sampler going.

With a draw of $\mathbf{U}_{\mathbf{ik}}^{*} \mid \boldsymbol{\beta}, \boldsymbol{\Sigma}^{*}, y_i, \mathbf{X}_{\mathbf{i}}; R_k$ in hand, we can then re-construct $\mathbf{U}_{\mathbf{i}}^{*}$ using

$$\mathbf{U}_{\mathbf{i}}^{*} = \left(\mathbf{D}_{\mathbf{k}}\right)^{-1} \mathbf{U}_{\mathbf{ik}}^{*} \tag{27}$$

The "Gibbs-within-Gibbs" Sampler is canned in function `GwG`, and implemented explicitly in function `gs_MNP_I`.

Multinomial Probit Model with identity COV Matrix

This is a simplified version with $\boldsymbol{\Sigma}$ set to the identity matrix. This makes sense when (i) choice sets vary over individuals, such that "option 1" etc. has no specific and invariant meaning, as would be the case in hypothetical choice experiments, OR (ii) when we have invariant choices (bus, taxi, train), and there is no policy plan to add another choice in the future, such that the IIA argument becomes moot.

IF the IIA dilemma poses a problem because additional options are considered in the future, error correlation over choices can be re-introduces via hierarchical distributions for model coefficients. Here, however, we describe the simplest case.

$$U_{ij} = \mathbf{x}_{\mathbf{ij}}'\boldsymbol{\beta} + \varepsilon_{ij} \qquad j = 1 \cdots J \qquad i = 1 \cdots N \tag{1}$$

$$\mathbf{U}_{\mathbf{i}} = \mathbf{X}_{\mathbf{i}}\boldsymbol{\beta} + \boldsymbol{\varepsilon}_{\mathbf{i}} \qquad \boldsymbol{\varepsilon}_{\mathbf{i}} \sim n\left(\mathbf{0}, \mathbf{I}\right) \tag{2}$$

Assume $\mathbf{x}_{\mathbf{i1}} = \mathbf{0}$. This implies

$$U_{i1} = \varepsilon_{ij} \tag{3}$$

i.e. the utility for the baseline alternative is completely random. Given the simplified structure of $\boldsymbol{\Sigma}$ we don't need to difference against the baseline or adjust for scale to identify all model parameters.

The latent model for the entire sample emerges as

$$\mathbf{U} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

$$\mathbf{U} = \underset{(n*J)x1}{\begin{bmatrix} \mathbf{U_1} \\ \mathbf{U_2} \\ \vdots \\ \mathbf{U_n} \end{bmatrix}} \qquad \mathbf{X} = \underset{(n*J)xk}{\begin{bmatrix} \mathbf{X_1} \\ \mathbf{X_2} \\ \vdots \\ \mathbf{X_n} \end{bmatrix}} \qquad \boldsymbol{\varepsilon} = \begin{bmatrix} \boldsymbol{\varepsilon_1} \\ \boldsymbol{\varepsilon_2} \\ \vdots \\ \boldsymbol{\varepsilon_n} \end{bmatrix} \sim n(\mathbf{0}, \boldsymbol{\Omega}) \qquad \text{with} \tag{4}$$

$$\boldsymbol{\Omega} = \underset{(n*J)x(n*J)}{\mathbf{I}}$$

$$y_i = k \quad \text{if} \qquad \max\left\{\left\{U_{ij}^*\right\}_{j=1}^{J}\right\} = U_{ik} \tag{5}$$

An individual's contribution to the likelihood function can thus be expressed as a multivariate normal *cdf*, with choice-specific truncation bounds. A choice of "*k*" implies:

$$\Pr(y_i = k) = \Pr\begin{pmatrix} -\infty < U_{i1}^* - U_{ik}^* \le 0 \\ -\infty < U_{i2}^* - U_{ik}^* \le 0 \\ \vdots \\ -\infty < U_{iJ}^* - U_{ik}^* \le 0 \end{pmatrix} = \Pr\begin{pmatrix} -\infty < (\varepsilon_{i1} - \varepsilon_{ki}) \le -\mathbf{x'_{ik}}\boldsymbol{\beta} \\ -\infty < (\varepsilon_{i2} - \varepsilon_{ki}) \le (\mathbf{x'_{i2}} - \mathbf{x'_{ik}})\boldsymbol{\beta} \\ \vdots \\ (\varepsilon_{iJ} - \varepsilon_{ki}) \le (\mathbf{x'_{iJ}} - \mathbf{x'_{ik}})\boldsymbol{\beta} \end{pmatrix} = \Phi(\mathbf{0}, \mathbf{D_k}\mathbf{D'_k}; R_k) \tag{6}$$

where (with slight abuse of notation) in this case $\Phi(.)$ denotes the *cdf* of the truncated multivariate normal density with mean $\mathbf{0}$ and variance matrix $\mathbf{D_k}\mathbf{D'_k}$, and truncation region R_j implicitly defined by the condition $(y_i = j)$. $\mathbf{D_k}$ is a (J-1) by J differencing matrix.

Generically, the likelihood contribution for the *ith* individual can be expressed as

$$p(y_i \mid \mathbf{X_i}, \boldsymbol{\beta}, \boldsymbol{\Sigma}) = \sum_{j=1}^{J}\left(\Phi(\mathbf{0}, \mathbf{D_j}\mathbf{D'_j}; R_j) I(y_i = j)\right) \tag{7}$$

leading to the sample likelihood function

$$p(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\Sigma}) = \prod_{i=1}^{n}\left(\sum_{j=1}^{J}\left(\Phi(\mathbf{0}, \mathbf{D_j}\mathbf{D'_j}; R_j) I(y_i = j)\right)\right) \tag{8}$$

In MLE, we would now use GHK or equivalent to evaluate the truncated *cdf* terms.
In Bayesian analysis we'll assign the usual normal prior to $\boldsymbol{\beta}$

$$p(\boldsymbol{\beta}) = (2\pi)^{-k/2} |\mathbf{V_0}|^{-1/2} \exp\left(-\tfrac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\mu_0})' \mathbf{V_0^{-1}} (\boldsymbol{\beta} - \boldsymbol{\mu_0})\right) \tag{9}$$

The augmented joint posterior takes the generic form of

$$p(\boldsymbol{\beta},\mathbf{U}\,|\,\mathbf{y},\mathbf{X}) \propto p(\boldsymbol{\beta})\,p(\mathbf{U}\,|\,\boldsymbol{\beta},\mathbf{X})\,p(\mathbf{y}\,|\,\boldsymbol{\beta},\mathbf{X},\mathbf{U}) \tag{10}$$

The first augmentation term, $p(\mathbf{U}\,|\,\boldsymbol{\beta},\mathbf{X})$, is a simple product of individual-specific multivariate normal density terms similar to the SUR model, i.e.

$$\begin{aligned} &p(\mathbf{U}\,|\,\boldsymbol{\beta},\mathbf{X}) = \prod_{i=1}^{n}(2\pi)^{-1}\exp\left(-\tfrac{1}{2}\left(\mathbf{U_i}-\mathbf{X_i}\boldsymbol{\beta}\right)'\left(\mathbf{U_i}-\mathbf{X_i}\boldsymbol{\beta}\right)\right) = \\ &(2\pi)^{-n}\exp\left(-\tfrac{1}{2}\sum_{i=1}^{n}\left(\mathbf{U_i}-\mathbf{X_i}\boldsymbol{\beta}\right)'\left(\mathbf{U_i}-\mathbf{X_i}\boldsymbol{\beta}\right)\right) \end{aligned} \tag{11}$$

The second term of the augmented data density can be written as

$$p(\mathbf{y}\,|\,\boldsymbol{\beta},\mathbf{X},\mathbf{U}) = p(\mathbf{y}\,|\,\mathbf{U}) = \prod_{i=1}^{n}\left(\sum_{k=1}^{J} I(y_i=k)\,I\left(\max\left\{\{U_{ij}\}_{j=1}^{J}\right\}=U_{ik}\right)\right)$$

The augmented joint posterior can now be explicitly written as

$$\begin{aligned} &p(\boldsymbol{\beta},\mathbf{U}\,|\,\mathbf{y},\mathbf{X}) \propto \\ &\exp\left(-\tfrac{1}{2}\left(\left(\boldsymbol{\beta}-\boldsymbol{\mu_0}\right)'\mathbf{V_0^{-1}}\left(\boldsymbol{\beta}-\boldsymbol{\mu_0}\right)\right)\right)*(2\pi)^{-n}\exp\left(-\tfrac{1}{2}\sum_{i=1}^{n}\left(\mathbf{U_i}-\mathbf{X_i}\boldsymbol{\beta}\right)'\left(\mathbf{U_i}-\mathbf{X_i}\boldsymbol{\beta}\right)\right)* \\ &\prod_{i=1}^{n}\left(\sum_{k=1}^{J} I(y_i=k)\,I\left(U_{ik}>\max\{\mathbf{U_{i-k}}\}\right)\right) \end{aligned}$$

The conditional posterior kernel for $\boldsymbol{\beta}$ now takes the following form:

$$p(\boldsymbol{\beta}\,|\,\mathbf{U},\mathbf{X}) \propto \exp\left(-\tfrac{1}{2}\left(\left(\boldsymbol{\beta}-\boldsymbol{\mu_0}\right)'\mathbf{V_0^{-1}}\left(\boldsymbol{\beta}-\boldsymbol{\mu_0}\right)+\sum_{i=1}^{n}\left(\mathbf{U_i}-\mathbf{X_i}\boldsymbol{\beta}\right)'\left(\mathbf{U_i}-\mathbf{X_i}\boldsymbol{\beta}\right)\right)\right) \tag{12}$$

This is again equivalent to the conditional posterior for the basic SUR model, and we can immediately derive the conditional posterior moments as:

$$\begin{aligned} &\boldsymbol{\beta}\,|\,\mathbf{U},\mathbf{X} \sim n(\boldsymbol{\mu_1},\mathbf{V_1}) \qquad \text{with} \qquad \mathbf{V_1} = \left(\mathbf{V_0^{-1}}+\sum_{i=1}^{n}\mathbf{X_i'X_i}\right)^{-1} \text{ and} \\ &\boldsymbol{\mu_1} = \mathbf{V_1}\left(\mathbf{V_0^{-1}}\boldsymbol{\mu_0}+\sum_{i=1}^{n}\mathbf{X_i'U_i}\right) \end{aligned} \tag{13}$$

For a given individual, the conditional posterior for latent utility takes the form of

$$p\left(\mathbf{U_i} \mid \boldsymbol{\beta}, y_i, \mathbf{X_i}\right) \propto$$
$$\exp\left(-\tfrac{1}{2}\left(\mathbf{U_i} - \mathbf{X_i}\boldsymbol{\beta}\right)'\left(\mathbf{U_i} - \mathbf{X_i}\boldsymbol{\beta}\right)\right) * \left(\sum_{k=1}^{J} I\left(U_{ik} > \max\left\{\mathbf{U_{i\text{-}k}}\right\}\right)\right) \tag{14}$$

This essentially denotes a multivariate normal density with mean $\mathbf{X_i}\boldsymbol{\beta}$ and identity variance, truncated to a specific region defined by the relative magnitude of the utility terms, as shown in (6).

An observed choice of *k* implies:

$$\Pr\left(y_i = k\right) = \Pr\begin{bmatrix} U_{i1} < U_{ik} \\ U_{i2} < U_{ik} \\ \vdots \\ U_{iJ} < U_{ik} \end{bmatrix} = \begin{bmatrix} 0 < U_{ik} - U_{i1} < \infty \\ 0 < U_{ik} - U_{i2} < \infty \\ \vdots \\ -\infty < U_{ik} - U_{iJ} < \infty \end{bmatrix} \tag{15}$$

So we can start by drawing $U_{ik} \mid \mathbf{U_{i,-k}}$ from its conditional univariate density (untruncated). Then we draw the remaining utilities from $U_{ij} \mid \mathbf{U_{i,-j}}$, $j \neq k$ form their conditional univariate density, truncated from above by U_{ik}.

Alternatively, we can work with differences as shown in the second matrix in (15), by specifying a J by J differencing matrix $\mathbf{D_k}$, which is $\mathbf{-I_k}$, with the *kth* column replaced by "1"s. We then draw $\mathbf{U_{ik}} = \mathbf{D_k}\mathbf{U_i}$ using adjusted means and variances.

After that, we can either re-construct $\mathbf{U_i} = \mathbf{D_k^{-1}}\mathbf{U_{ik}}$ and draw $\boldsymbol{\beta} \mid \mathbf{U}, \mathbf{X}$ as outlined in (12) and (13), or keep utilities and $\mathbf{X}$ in differenced form and draw the coefficients given the differenced utility and data. The latter is the strategy taken in the Gibbs sampler for this example, i.e. `gs_MNP_I`.

References:

Geweke, J. 1991. "Efficient Simulation from the Multivariate Normal and Student-t Distributions Subject to Linear Constraints and the Evaluation of Constraint Probabilities," Working paper, Department of Economics, University of Minnesota.

Hausman, J. and Wise, D. 1978. " A Conditional Probit Model for Qualitative Choice: Discrete Decisions Recognizing Interdependence and Heterogeneous Preferences," *Econometrica*, 46: 403-426.

Layton, D. F., and R. A. Levine. 2005. "Bayesian Approaches to Modeling Stated Preference Data." in R. Scarpa and A. Alberini, eds. *Applications of Simulation Methods in Environmental and Resource Economics*. Dordrecht, The Netherlands: Springer, pp. 187-205.

Layton, D. F., and R. A. Levine. 2003. "How Much Does the Far Future Matter? A Hierarchical Bayesian Analysis of the Public's Willingness to Mitigate Ecological Impacts of Climate Change." *Journal of the American Statistical Association* 98: 533-544.

Moeltner K., R. Johnston, R. S. Rosenberger, J. M. Duke. 2009. "Benefit Transfer from Multiple Contingent Experiments: A Flexible Two-Step Model Combining Individual Choice Data with Community Characteristics," *American Journal of Agricultural Economics*, 91: 1335-42.

Robert, C. P. 1995. "Simulation of Truncated Normal Variables." *Statistics and Computing* 5: 121-125.

Train, K. 2003. *Discrete choice methods with simulation*. Cambridge University Press