Bayesian inference for factor structure models via gibbs sampling

Piatek, Rémi

CENTER FOR QUANTITATIVE METHODS AND SURVEY RESEARCH

CMS Discussion Paper 2010-03

Bayesian Inference for Factor Structure Models via Gibbs Sampling

Rémi Piatek*

*University Konstanz, Department of Economics

http://www.uni-konstanz.de/cms

Bayesian Inference for Factor Structure Models via Gibbs Sampling

Rémi Piatek

Department of Economics, University of Konstanz, Box D 124, 78457 Konstanz, Germany (e-mail: remi.piatek@uni-konstanz.de)

Abstract

The goal of this paper is to provide all the technical details required to implement Gibbs sampling for the estimation of simultaneous equation models with common latent factors among their regressors, so-called 'factor structure models.' Linear, dichotomous and censored response models, as well as ordered and unordered response models can be accommodated in this framework. The latent factors can be either correlated or not, and specified as normally distributed or as following a finite mixture of normal distributions for more flexibility. All conditional distributions are derived and can be used to construct the Gibbs sampler step by step.

JEL classification: C11; C31.
Keywords: Latent factor models; Simultaneous equation models; Markov Chain Monte Carlo; Normal Mixture.

1. INTRODUCTION

Factor structure models have become a very popular tool in the social sciences, and especially in economics. This appeal is mainly due to the fact that these models, by combining the advantages of simultaneous equation models and factor analytic models, allow the econometrician to tackle some pervasive problems such as measurement error and endogeneity. For instance, they make it possible to investigate the impact of latent abilities or personality traits measured with error by some indicators on various economic and social outcomes (for recent examples, see Hansen et al. 2004; Heckman et al. 2006).

Among all the statistical methods available for the estimation of this type of model, Bayesian methods have been extensively used by some authors based on their convenience (Carneiro et al. 2003). They appear to be particularly well suited to such models where the number of equations and parameters can be very large. However, the features of the Gibbs sampler have not been documented in a comprehensive way for this kind of problem. Empirical researchers who are unfamiliar with Bayesian methods may therefore become discouraged and give up using this estimation strategy, or even worse give up using factor structure models altogether. This paper aims to fill this gap by presenting all the technical details required to construct the Gibbs sampler step by step.

1.1. Model specification

Throughout this article, we will try to remain as general as possible with respect to the framework of the model, so that many different models can easily be accommodated. The overall model consists of S submodels and can be expressed as a set of S equations of the following form:

$$\begin{aligned} Y_{s,i} &= g_s(Y^*_{s,i}; \sigma_s), & s &= 1, ..., S, \\ Y^*_{s,i} &= X_{s,i}\beta_s + \alpha_s \Theta_i + \varepsilon_{s,i}, & i &= 1, ..., n, \end{aligned}$$

where $Y_{s,i}$ is the observed outcome s for individual i which results from the transformation of the corresponding latent outcome $Y^*_{s,i}$ by the link function g_s. This function is parameterized by some parameters σ_s and the observed outcome can be continuous, binary, categorical or ordinal.

The vector Θ_i contains J single latent factors which influence the outcome $Y_{s,i}$ trough factor loadings contained in the $(1 \times J)$-dimensional vector α_s. Usually, the factors are assumed to be independent of the covariates and of the error terms $(\Theta \perp\!\!\!\perp X, \varepsilon)$ for identification purposes. The parameters defining the distribution of the latent factors are denoted ψ_Θ. Specification of the latent factors is very flexible: they can be correlated or not, and normally distributed or specified as following a mixture of normals for more flexibility.

The outcome is also affected by some covariates X_s through the slope parameters β_s. Covariates can be different in each submodel, but for the sake of convenience they will be denoted by X without distinction in the following. If an intercept term is required, a vector of 1's has to be included among the regressors X.

The error term ε_s is an i.i.d. random variable associated with the outcome s which is usually specified as normally distributed, or following a mixture of normals for more flexibility in some cases. Error terms are independent across equations. Let $\psi_{\varepsilon,s}$ be the set of parameters defining the distribution of ε_s.[1]

This paper will not explicitly deal with identification. Identification is a typical problem in factor analysis and is now well-documented. For instance, Carneiro et al. (2003) and Hansen et al. (2004) demonstrate and discuss the identification of factor structure models.

1.2. Likelihood and posterior

Let Y and Y^* be, respectively, the matrices of all observed and latent outcomes in the overall model. Let $\delta_s = (\beta_s, \alpha_s, \psi_{\varepsilon,s}, \sigma_s)$ denote the parameters specific to model s, and $\delta = (\delta_1, ..., \delta_S, \psi_\Theta)$ the parameters of the overall model. The likelihood can be written as:

$$\begin{aligned} \mathcal{L}(\delta|Y, X) &= \iint f(Y, Y^*, \Theta|X, \delta) dY^* d\Theta, \\ &= \iint f(Y|Y^*, \Theta, X, \delta) f(Y^*|\Theta, X, \delta) f(\Theta|\delta) dY^* d\Theta. \end{aligned}$$

Because of the common factors Θ across equations, the observed outcomes are not independent and deriving a closed-form expression for the likelihood appears to be cumbersome in most cases. To circumvent this problem, data augmentation procedures (Tanner and Wong

[1] Precision in the normal case; component weights, means and precisions in the mixture case.

1987) can be implemented to make the sampling procedure easier.[2] The latent variables—latent outcomes and factors—are treated as unobservables and are explicitly introduced in the likelihood function. Applying Bayes' rule, the posterior is then derived as follows:

$$\begin{aligned} f(\delta, Y^*, \Theta|Y, X) &\propto f(Y, Y^*, \Theta|X, \delta) f(\delta), \\ &\propto f(Y|Y^*, \sigma) f(Y^*|\Theta, X, \delta) f(\Theta|\delta) f(\delta), \\ &\propto \prod_{s=1}^{S} f(Y_s|Y_s^*, \sigma_s) f(Y_s^*|\Theta, X, \delta_s) f(\delta_s) f(\Theta|\delta). \end{aligned}$$

The latent factors Θ are the only source of dependence across the S equations. From the last equation, it appears that conditional on Θ, the different submodels are independent of each other and can then be treated separately. This is one of the main advantages of data augmentation in this context. Once the latent factors have been simulated, the whole problem can be divided into as many distinct tasks as there are different submodels. The Gibbs sampler is therefore particularly appropriate for this kind of problem (Casella and George 1992).

1.3. Set-up of the Gibbs sampler

The algorithm is initialized by choosing initial values for all parameters δ, all latent outcomes Y^* and all latent factors Θ. Random values can be chosen, or parameter estimates from some preliminary analysis.[3] Once the initialization has been achieved, the Gibbs sampling scheme is implemented sequentially. At each iteration (t):

1. For each submodel s, update the parameters δ_s and the latent outcomes given the latent factors of the previous iteration:
   (a) draw the slope parameters $\beta_s^{(t)}$ from $f(\beta_s|Y_s^{*(t-1)}, X, \alpha_s^{(t-1)}, \Theta^{(t-1)}, \psi_{\varepsilon,s}^{(t-1)})$,
   (b) draw the factor loadings $\alpha_s^{(t)}$ from $f(\alpha_s|Y_s^{*(t-1)}, X, \beta_s^{(t)}, \Theta^{(t-1)}, \psi_{\varepsilon,s}^{(t-1)})$,
   (c) draw the latent outcomes $Y_{s,i}^{*(t)}$ from $f(Y_{s,i}^*|X_i, \beta_s^{(t)}, \alpha_s^{(t)}, \Theta_i^{(t-1)}, \psi_{\varepsilon,s}^{(t-1)})$ for each individual i,
   (d) draw the model-specific parameters $\sigma_s^{(t)}$, if any, from $f(\sigma_s|Y_s, Y_s^{*(t)})$.
   (e) draw the error term parameters $\psi_{\varepsilon,s}^{(t)}$ from $f(\psi_{\varepsilon,s}|Y_s^{*(t)}, X, \beta_s^{(t)}, \alpha_s^{(t)}, \Theta^{(t-1)})$,
2. Update the factors. Draw $\Theta_i^{(t)}$ from $f(\Theta_i|Y_i^{*(t)}, X_i, \beta^{(t)}, \alpha^{(t)}, \psi_\varepsilon^{(t)}, \psi_\Theta^{(t-1)})$ for each individual i.
3. Draw the parameters of the distribution of the factors $\psi_\Theta^{(t)}$ from $f(\psi_\Theta|\Theta^{(t)})$.
4. Go to step 1 and repeat until practical convergence.

Step 1 is described in Sections 2–6 for each type of submodel—linear, dichotomous, censored, categorical and ordinal. Because of the data augmentation scheme, the procedure used to sample the slope parameters, the factor loadings and the error term parameters is the same for the submodels involving latent outcomes as found in the linear case. Once the latent

[2] See also van Dyk and Meng (2001) for a review.
[3] E.g., exploratory factor analysis for the starting values of the factor loadings.

outcomes have been updated, they can actually be treated as given, and the linear model can therefore be used as a baseline. For this reason, the linear case will first be thoroughly developed and the demonstration will not be repeated for the subsequent submodels. Only the conditional distributions of the latent outcomes and of the model-specific parameters such as cut-points will then be provided. Section 7 describes how to sample the latent factors conditional on all the other parameters in step 2, and Section 8 then explains how to sample the parameters of the distribution of the factors conditional on the factors in step 3. Section 9 presents the procedure used to update the parameters of a mixture of normal distributions implemented in the error terms and the latent factors in the mixed case. Section 10 finally concludes.

To simplify the notation, no differences will be made between the different submodels and henceforth the subscript s will be dropped. In the same way, the parameter superscripts referring to the current and previous iterations will not be mentioned anymore. Obviously, conditional distributions which depend on other parameters use the latest updated values of these parameters.

2. LINEAR RESPONSE SUBMODELS

In the linear case, the link function g is the identity function and the model is:

$$Y_i = X_i\beta + \alpha\Theta_i + \varepsilon_i.$$

2.1. Updating the slope parameters

Applying Bayes' rule, the conditional distribution of β is derived as:

$$f(\beta|Y, X, \Theta, \alpha, \psi_\varepsilon) \propto f(Y|X, \Theta, \beta, \alpha, \psi_\varepsilon) f(\beta). \tag{1}$$

A conjugate normal prior centered at μ_β and with precision matrix Ψ_β is assumed for the slopes:

$$\beta \sim \mathcal{N}(\mu_\beta\,;\,\Psi_\beta)\,.$$

Using the auxiliary outcome $\widetilde{Y} = Y - \alpha\Theta$, Equation (1) implies that:

$$\begin{aligned}
&f(\beta|Y, X, \Theta, \alpha, \psi_\varepsilon) = f(\beta|\widetilde{Y}, X, \psi_\varepsilon),\\
&\propto f(\widetilde{Y}|X, \beta, \tau_\varepsilon) f(\beta),\\
&\propto \exp\left\{-\frac{\tau_\varepsilon}{2}(\widetilde{Y} - X\beta)'(\widetilde{Y} - X\beta)\right\} \exp\left\{-\frac{1}{2}(\beta - \mu_\beta)'\Psi_\beta(\beta - \mu_\beta)\right\},\\
&\propto \exp\left\{-\frac{1}{2}\left(\beta'\left[\tau_\varepsilon X'X + \Psi_\beta\right]\beta - 2\beta'\left[\tau_\varepsilon X'\widetilde{Y} + \Psi_\beta\mu_\beta\right]\right)\right\},\\
&\propto \exp\left\{-\frac{1}{2}\left(\beta - \left[\tau_\varepsilon X'X + \Psi_\beta\right]^{-1}\left[\tau_\varepsilon X'\widetilde{Y} + \Psi_\beta\mu_\beta\right]\right)'\left(\tau_\varepsilon X'X + \Psi_\beta\right)(\bullet)\right\},
\end{aligned}$$

where $(\bullet)$ represents the first factor of the corresponding sandwich matrix. Factors not involving β have been omitted, and the resulting normal kernel has been produced using the completion of the square. As a consequence, β has the following conditional distribution:

$$\beta \sim \mathcal{N}\Big([\tau_\varepsilon X'X + \Psi_\beta]^{-1}[\tau_\varepsilon X'\widetilde{Y} + \Psi_\beta\mu_\beta]\,;\,\tau_\varepsilon X'X + \Psi_\beta\Big)\,.$$

When noninformative priors are assumed ($\mu_\beta = \{0\}$ and $\Psi_\beta = \{0\}$), the conditional distribution of β reduces to:

$$\beta \sim \mathcal{N}\Big([X'X]^{-1}X'\widetilde{Y}\,;\,\tau_\varepsilon X'X\Big)\,.$$

2.2. Updating the factor loadings

Applying Bayes' rule for α provides:

$$f(\alpha|Y,X,\Theta,\beta,\psi_\varepsilon) \propto f(Y|X,\Theta,\beta,\alpha,\psi_\varepsilon)f(\alpha).$$

Since the factors Θ are updated at each iteration of the Gibbs sampler, they can be regarded as given when it comes to sampling the factor loadings. Using their current values, the factor loadings are updated with exactly the same sampling scheme as for the slope parameters.

2.2.1 Unconstrained factor loadings

Factor loadings are supposed to be a priori normally distributed:

$$\alpha \sim \mathcal{N}(\mu_\alpha\,;\,\Psi_\alpha)\,.$$

In the same way as we did for the slope parameters, we can show that the conditional of the loadings is the following normal distribution:

$$\alpha \sim \mathcal{N}\Big([\tau_\varepsilon\Theta'\Theta + \Psi_\alpha]^{-1}[\tau_\varepsilon\Theta'\widetilde{Y} + \Psi_\alpha\mu_\alpha]\,;\,\tau_\varepsilon\Theta'\Theta + \Psi_\alpha\Big)\,,$$

where $\widetilde{Y} = Y - X\beta$.

In many applications, a prior with mean zero and with a given precision τ_α is used, yielding the following conditional distribution:

$$\alpha \sim \mathcal{N}\Big([\tau_\varepsilon\Theta'\Theta + \tau_\alpha I_J]^{-1}\tau_\varepsilon\Theta'\widetilde{Y}\,;\,\tau_\varepsilon\Theta'\Theta + \tau_\alpha I_J\Big)\,,$$

where I_J is the identity matrix of dimension J.

2.2.2 Constrained factor loadings

It might be desired—or required—to restrict some factor loadings in some submodels. In the case where a factor loading has to be set to a given value, it is simply assigned this value at each step of the Gibbs sampler.[4] If sign constraints have to be implemented, the sampling procedure is very similar to the unrestricted case, with the only difference that the conditional is a truncated normal distribution:

$$\alpha \sim \mathcal{TN}_{(0,+\infty)}\Big([\tau_\varepsilon\Theta'\Theta + \tau_\alpha Id]^{-1}\tau_\varepsilon\Theta'\widetilde{Y}\,;\,\tau_\varepsilon\Theta'\Theta + \tau_\alpha Id\Big) \quad \text{for positive factor loadings,}$$

$$\alpha \sim \mathcal{TN}_{(-\infty,0]}\Big([\tau_\varepsilon\Theta'\Theta + \tau_\alpha Id]^{-1}\tau_\varepsilon\Theta'\widetilde{Y}\,;\,\tau_\varepsilon\Theta'\Theta + \tau_\alpha Id\Big) \quad \text{for negative factor loadings,}$$

where in this case $\widetilde{Y} = Y - X\beta - \Theta^f\alpha^f$, with Θ^f the set of unrestricted factors and α^f their corresponding loadings.

[4] In factor analysis, it is often required that the scale of the latent factor be set by, for instance, fixing one of the loadings to 1.

2.3. Updating error term parameters

The application of Bayes' rule for ψ_ε implies that:

$$f(\psi_\varepsilon|Y,X,\Theta,\beta,\alpha) \propto f(Y|X,\Theta,\beta,\alpha,\psi_\varepsilon)f(\psi_\varepsilon).$$

Depending on the type of distribution assumed for the error term, two different cases have to be distinguished.

2.3.1 *Normal error term* $\varepsilon \sim \mathcal{N}(0\,;\,\tau_\varepsilon)$

In the normal case, a single parameter defines the distribution of the error term ($\psi_\varepsilon = \tau_\varepsilon$). Because the precision has to be positive, a conjugate Gamma distribution prior is usually used:

$$\tau_\varepsilon \sim \mathcal{G}a(g_1\,;\,g_2)\,,$$

and the conditional is proportional to:

$$\begin{aligned} f(\tau_\varepsilon|Y,X,\Theta,\beta,\alpha) &\propto f(Y|X,\Theta,\beta,\alpha,\tau_\varepsilon)f(\tau_\varepsilon), \\ &\propto \tau_\varepsilon^{n/2}\exp\left\{-\frac{\tau_\varepsilon}{2}\sum_{i=1}^{n}\left(Y_i - X_i\beta - \alpha\Theta_i\right)^2\right\}\tau_\varepsilon^{g_1-1}\exp\left\{-g_2\tau_\varepsilon\right\}, \\ &\propto \tau_\varepsilon^{n/2+g_1-1}\exp\left\{-\tau_\varepsilon\left(g_2 + \frac{1}{2}\sum_{i=1}^{n}\left(Y_i - X_i\beta - \alpha\Theta_i\right)^2\right)\right\}. \end{aligned}$$

The kernel of a Gamma distribution can be recognized in this last expression. Hence, τ_ε has the following conditional distribution:

$$\tau_\varepsilon \sim \mathcal{G}a\left(g_1 + \frac{n}{2}\,;\,g_2 + \frac{1}{2}\sum_{i=1}^{n}\left(Y_i - X_i\beta - \alpha\Theta_i\right)^2\right).$$

2.3.2 *Mixed error term*

A finite mixture of normals can be assumed for the distribution of the error term, in order to introduce more flexibility. This boils down to considering that individuals belong to different subgroups with homogenous non-observable characteristics.

In this context, the error term is assumed to be generated by a mixture of K normals, with mean constrained to zero for the purpose of identification:

$$\varepsilon \sim \sum_{k=1}^{K} p_{\varepsilon,k}\mathcal{N}(\mu_{\varepsilon,k}\,;\,\tau_{\varepsilon,k})\,, \qquad\qquad E(\varepsilon) = \sum_{k=1}^{K} p_{\varepsilon,k}\mu_{\varepsilon,k} = 0,$$

where $p_{\varepsilon,k}$ is the weight, $\mu_{\varepsilon,k}$ is the mean and $\tau_{\varepsilon,k}$ is the precision of mixture component k.

The complete procedure for updating the parameters of a mixture of normals with constrained mean is described in Section 9.

The introduction of a mixture of normals for the error term makes the initial problem slightly more complicated, insofar as the conditional distribution of the outcome is also a

mixture of normals under this assumption. However, a simple transformation of the variables solves the problem. Once the mixture parameters have been updated, the dependent and independent variables can be redefined as follows to take into account the heterogeneity of the error term:

$$\begin{aligned} Y_i^{new} &\leftarrow \tau_{\varepsilon,k}^{1/2}(Y_i - \mu_{\varepsilon,k}), \\ X_i^{new} &\leftarrow \tau_{\varepsilon,k}^{1/2} X_i, \\ \Theta_i^{new} &\leftarrow \tau_{\varepsilon,k}^{1/2} \Theta_i, \end{aligned}$$

where individual i belongs to mixture group k. After this transformation, the error term appears to be standard normally distributed for all individuals. It is therefore enough to use these transformed variables instead of the original ones, and to set the precision τ_ε to unity when deriving the conditional distributions of the other parameters.

3. DICHOTOMOUS RESPONSE SUBMODELS

In this type of submodel, the link function is the indicator function $\mathbb{1}[\cdot]$ and the outcome is a binary variable defined by the sign of its underlying latent outcome Y^*:

$$\begin{aligned} Y_i &= \mathbb{1}[Y_i^* > 0], \\ Y_i^* &= X_i\beta + \alpha\Theta_i + \varepsilon_i. \end{aligned}$$

Conditional on the latent factor Θ, this is the well-known probit model when $\varepsilon \sim \mathcal{N}(0\,;\,1)$.

The conditional of the parameters and of the latent outcome Y^* can be factorized as follows:

$$f(\beta, \alpha, Y^* | Y, X, \Theta) \propto f(Y|Y^*) f(Y^*|X, \Theta, \beta, \alpha) f(\beta, \alpha). \tag{2}$$

3.1. Sampling the latent outcome

Assuming that $\varepsilon \sim \mathcal{N}(0\,;\,1)$, the conditional of the latent outcome can be derived from Equation (2) as follows:

$$\begin{aligned} f(Y^*|Y, X, \Theta, \beta, \alpha) &\propto f(Y|Y^*) f(Y^*|X, \Theta, \beta, \alpha), \\ &\propto \prod_{i=1}^{n} \mathbb{1}[Y_i^* > 0]^{Y_i} \, \mathbb{1}[Y_i^* \leqslant 0]^{1-Y_i} \prod_{i=1}^{n} \exp\left\{-\frac{1}{2}(Y_i^* - X_i\beta - \alpha\Theta_i)^2\right\}, \\ &\propto \prod_{i=1}^{n} \left[\mathbb{1}[Y_i^* > 0] \exp\left\{-\frac{1}{2}(Y_i^* - X_i\beta - \alpha\Theta_i)^2\right\}\right]^{Y_i} \\ &\quad \times \left[\mathbb{1}[Y_i^* \leqslant 0] \exp\left\{-\frac{1}{2}(Y_i^* - X_i\beta - \alpha\Theta_i)^2\right\}\right]^{1-Y_i}. \end{aligned}$$

Hence, the latent outcome will be sampled from a truncated normal distribution, dependent on the outcome of the observed binary variable:

$$Y_i^* \sim \begin{cases} \mathcal{TN}_{(0,+\infty)}(X_i\beta + \alpha\Theta_i\,;\,1) & \text{if } Y_i = 1, \\ \mathcal{TN}_{(-\infty,0]}(X_i\beta + \alpha\Theta_i\,;\,1) & \text{if } Y_i = 0. \end{cases}$$

In case the error term follows a mixture of normal distributions, the conditional of the latent outcome is:

$$Y_i^* \sim \begin{cases} \mathcal{TN}_{(0,+\infty)}(X_i\beta + \alpha\Theta_i + \mu_{\varepsilon,k}\,;\,\tau_{\varepsilon,k}) & \text{if } Y_i = 1, \\ \mathcal{TN}_{(-\infty,0]}(X_i\beta + \alpha\Theta_i + \mu_{\varepsilon,k}\,;\,\tau_{\varepsilon,k}) & \text{if } Y_i = 0, \end{cases}$$

where $\mu_{\varepsilon,k}$ and $\tau_{\varepsilon,k}$ are, respectively, the mean and the precision of the error term component corresponding to the mixture group k individual i belongs to.

3.2. Updating the error term parameters

This step is only necessary when the standard normal assumption on the error term is relaxed, and a mixture of normals with zero mean is assumed instead—with same specification as in the linear case (see Section 2). Geweke and Keane (1999) have proposed the use of mixture of normals probit models as a flexible approach to dichotomous models. Because they remove the usual normality assumption, they are much closer to semiparametric models. Some normalizations are however required for identification. This can be achieved in different ways:

- scale mixture of normals models: $\mu_{\varepsilon,k} = 0 \quad (k = 1, ..., K)$,
- mean mixture of normals models: $\tau_{\varepsilon,k} = 1 \quad (k = 1, ..., K)$,
- full mixture of normals models:
  - $\tau_{\varepsilon,k} > 0 \quad \forall k$,
  - either $\mu_{\varepsilon,k-1} < \mu_{\varepsilon,k}$ or $\tau_{\varepsilon,k-1} < \tau_{\varepsilon,k} \quad (k = 2, ..., K)$,
  - $\tau_{\varepsilon,k} = 1$ for some k.

Further regularity assumptions are discussed in Geweke and Keane (1999). In this configuration, the mixture parameters are updated according to the procedure described in Section 9, using the latent outcomes Y^* as dependent variable. Like in the linear case, the latent outcome, the factors and the explanatory variables have to be redefined once the mixture parameters have been updated, to take into account the heterogeneity of the error term.

4. CENSORED RESPONSE SUBMODELS

The latent outcome of the censored submodel is observed only above a given threshold σ:[5]

$$\begin{aligned} Y_i &= Y_i^* \mathbb{1}\left[Y_i^* > \sigma\right], \\ Y_i^* &= X_i\beta + \alpha\Theta_i + \varepsilon_i. \end{aligned}$$

When $\varepsilon \sim \mathcal{N}(0\,;\,\tau_\varepsilon)$, this is the typical Tobit-I model, conditionally on Θ. Bayesian inference in this type of model is detailed in Chib (1992).

[5] The model can also be censored from above. In this case, the latent outcome is observed only if it is below the threshold σ and the modification of the following sampling procedure is straightforward.

4.1. Sampling the latent outcome

Only the case where the observed outcome is censored requires elaboration. For each individual i, the conditional density of the latent outcome given censoring can be expressed as:

$$\begin{aligned} f(Y_i^*|Y_i = 0, X_i, \sigma, \Theta_i, \beta, \alpha, \tau_\varepsilon) &\propto f(Y_i = 0|Y_i^*, \sigma) f(Y_i^*|X_i, \Theta, \beta, \alpha, \tau_\varepsilon), \\ &\propto \mathbb{1}\left[Y_i^* \leqslant \sigma\right] \exp\left\{-\frac{\tau_\varepsilon}{2}(Y_i^* - X_i\beta - \alpha\Theta_i)^2\right\}. \end{aligned}$$

Hence, the latent outcome is sampled from the following truncated normal distribution:[6]

$$Y_i^* \sim \mathcal{TN}_{(-\infty,\sigma]}(X_i\beta + \alpha\Theta_i\,;\,\tau_\varepsilon)\,.$$

Similarly, in the case where the error term follows a mixture of normals with zero mean—with same specification as in the dichotomous case—, the latent outcome is sampled from:

$$Y_i^* \sim \mathcal{TN}_{(-\infty,\sigma]}(X_i\beta + \alpha\Theta_i + \mu_{\varepsilon,k}\,;\,\tau_{\varepsilon,k})\,,$$

where $\mu_{\varepsilon,k}$ and $\tau_{\varepsilon,k}$ are, respectively, the mean and the precision of the error term component corresponding to the mixture group k individual i belongs to.

In the next steps of the Gibbs sampler, the simulated latent outcome is used for the individuals whose outcome is censored, while the observed outcome is used for the others.

5. UNORDERED RESPONSE SUBMODELS

Assume there are L different unordered choices and consider the following utility maximization problem:

$$\begin{aligned} Y_i &= \underset{l=1,...,L}{\text{argmax}}\; Y_{l,i}^*, \\ Y_{l,i}^* &= X_{l,i}\beta_l + \alpha_l\Theta_i + \varepsilon_{l,i}, \qquad\qquad l = 1, ..., L. \end{aligned}$$

Conditional on Θ, this is the standard multinomial probit model when ε_l is normally distributed.

A well-known problem of identification arises in this class of model (Bunch 1991). Although this issue is not addressed here, appropriate restrictions or modifications of the baseline model should be carefully implemented to secure identification (McCulloch and Rossi 2000; Keane 1992). Modification of the sampling procedure presented in the next section to fit the identification requirements would be straightforward. In the following, the error terms are assumed to be standard normally distributed.

5.1. Sampling the latent outcomes

Since there are L different choices and as many underlying equations, the easiest way to sample the latent outcomes is to do so sequentially. For individual i, the conditional distribution of her latent outcome $Y_{m,i}^*$ for choice m conditional on the other latent outcomes is:

$$f(Y_{m,i}^*|Y_i, Y_{-m,i}^*, X, \Theta, \beta, \alpha) \propto f(Y_i|Y_{m,i}^*, Y_{-m,i}^*) f(Y_{m,i}^*|X, \Theta, \beta, \alpha),$$

[6] When the outcome is censored from above, the truncation is on the interval $(\sigma, +\infty)$.

where $m = 1, ..., L$ and $Y^*_{-m,i}$ stands for the set of latent outcomes excluding the m^{th} one.

If m is the actual choice ($Y_i = m$), then the conditional is:

$$f(Y^*_{m,i}|Y_i = m, Y^*_{-m,i}, X, \Theta, \beta, \alpha) \propto \mathbb{1}\left[Y^*_{m,i} \geqslant \max_{\substack{l=1,...,L \\ l \neq m}} \{Y^*_{l,i}\}\right] \exp\left(-\frac{1}{2}(Y^*_{m,i} - X_{m,i}\beta_m - \alpha_m\Theta_i)^2\right), \tag{3}$$

and otherwise, if $Y_i \neq m$:

$$f(Y^*_{m,i}|Y_i \neq m, Y^*_{-m,i}, X, \Theta, \beta, \alpha) \propto \mathbb{1}\left[Y^*_{m,i} < Y^*_{Y_i,i}\right] \exp\left(-\frac{1}{2}(Y^*_{m,i} - X_{m,i}\beta_m - \alpha_m\Theta_i)^2\right). \tag{4}$$

From Equations (3) and (4), it can be deduced that each latent outcome $Y^*_{m,i}$ is sampled sequentially from a truncated normal distribution, dependent on the other latent outcomes and on the observed choice as follows:

$$Y^*_{m,i} \sim \begin{cases} \mathcal{TN}_{(\max_{l\neq m}\{Y^*_{l,i}\};+\infty)}(X_{m,i}\beta_m + \alpha_m\Theta_i\,;\,1) & \text{if } Y_i = m, \\ \mathcal{TN}_{(-\infty;Y^*_{Y_i,i})}(X_{m,i}\beta_m + \alpha_m\Theta_i\,;\,1) & \text{otherwise,} \end{cases}$$

for $m = 1, ..., L$.

6. ORDERED RESPONSE SUBMODELS

As in the unordered case, there are L different choices. But in this configuration, they have a natural ordering generated by a latent outcome Y^* which depends on some covariates and on the factors:

$$\begin{aligned} Y_i &= l \quad \text{if } c_{l-1} \leqslant Y^*_i < c_l, \qquad\qquad l = 1, ..., L, \\ Y^*_i &= X_i\beta + \alpha\Theta_i + \varepsilon_i, \end{aligned}$$

where $c_0 = -\infty$, $c_L = +\infty$ and $c = (c_1, ..., c_{L-1})'$. When the error term is standard normally distributed, this is the usual ordered probit model, conditionally on Θ.

The conditional distribution of the parameters can be factorized as follows:

$$f(\beta, \alpha, c|Y, Y^*, X, \Theta) \propto f(Y|Y^*, c)f(Y^*|X, \Theta, \beta, \alpha)f(\beta, \alpha)f(c). \tag{5}$$

Following Albert and Chib (1993), a data augmentation scheme is implemented to sample this model and the latent outcome Y^* is simulated at each step of the algorithm.

6.1. Sampling the latent outcome

Conditional on the observed choices Y_l, on the factor and on all parameters, the conditional distribution of the latent outcome is, when $\varepsilon_i \sim \mathcal{N}(0\,;\,1)$:

$$f(Y^*|Y, X, \Theta, \beta, \alpha, c) \propto f(Y|Y^*, c)f(Y^*|X, \Theta, \beta, \alpha),$$

$$\propto \prod_{i=1}^{n} \left\{ \sum_{l=1}^{L} \mathbb{1}\left[Y_i = l\right] \mathbb{1}\left[c_{l-1} \leqslant Y_i^* < c_l\right] \right\} \mathbb{1}\left[c_0 < ... < c_L\right]$$
$$\times \prod_{i=1}^{n} \exp\left(-\frac{1}{2}(Y_i^* - X_i\beta - \alpha\Theta_i)^2 \right),$$
$$\propto \prod_{i=1}^{n} \left\{ \sum_{l=1}^{L} \mathbb{1}\left[c_{l-1} \leqslant Y_i^* < c_l\right] \exp\left(-\frac{1}{2}(Y_i^* - X_i\beta - \alpha\Theta_i)^2 \right) \right\}.$$

Hence, Y^* is sampled from a truncated normal distribution:

$$Y_i^* \sim \mathcal{TN}_{[c_{l-1},c_l)}(X_i\beta + \alpha\Theta_i\,;\,1) \qquad \text{whenever } Y_i = l.$$

Similarly, it can be shown that when the error is mixed distributed, the conditional distribution of Y^* is:

$$Y_i^* \sim \mathcal{TN}_{[c_{l-1},c_l)}(X_i\beta + \alpha\Theta_i + \mu_{\varepsilon,k}\,;\,\tau_{\varepsilon,k}) \qquad \text{whenever } Y_i = l,$$

where $\mu_{\varepsilon,k}$ and $\tau_{\varepsilon,k}$ are, respectively, the mean and the precision of the error term component corresponding to the mixture group k individual i belongs to.[7]

6.2. Updating the cut-points

6.2.1 *Standard Gibbs sampler*

Albert and Chib (1993) have developed a simple procedure to sample the cut-points. Conditional on β and α, the factorization of Equation (5) provides:

$$f(c|Y,Y^*,X,\Theta,\beta,\alpha) \propto f(Y|Y^*,c)f(c),$$
$$\propto \prod_{i=1}^{n} \left\{ \sum_{l=1}^{L} \mathbb{1}\left[Y_i = l\right] \mathbb{1}\left[c_{l-1} \leqslant Y_i^* < c_l\right] \right\} \mathbb{1}\left[c_1 < ... < c_{L-1}\right] f(c).$$

The cut-points can be sampled sequentially. The conditional distribution of c_l is:

$$f(c_l|c_{-l},Y,Y^*,X,\Theta,\beta,\alpha) \propto \prod_{i:Y_i=l} \mathbb{1}\left[c_{l-1} \leqslant Y_i^* < c_l\right] \prod_{i:Y_i=l+1} \mathbb{1}\left[c_l \leqslant Y_i^* < c_{l+1}\right]$$
$$\times \mathbb{1}\left[c_{l-1} < c_l < c_{l+1}\right] f(c_l), \tag{6}$$

where c_{-l} is the set of cut-points other than c_l. For each cut-point c_l, a uniform prior on the interval bounded by c_{l-1} and c_{l+1} is usually used, so as to fulfill the ordering condition. From Equation (6), it can be deduced that the conditional of each cut-point is a uniform distribution:

$$c_l \sim U(\underline{c}_l;\overline{c}_l),$$

with:

$$\underline{c}_l = \max\left\{ \max_{i:Y_i=l}\left\{Y_i^*\right\}\,;\,c_{l-1} \right\} \qquad \text{and} \qquad \overline{c}_l = \min\left\{ \min_{i:Y_i=l+1}\left\{Y_i^*\right\}\,;\,c_{l+1} \right\}.$$

[7] The specification of the mixture is the same as in the dichotomous case (see Section 3).

The standard Gibbs sampler can exhibit low convergence rate, due to the high correlation between the latent outcomes and the cut-points. To remedy this problem, two different approaches can be implemented. The first one is based on the work by Cowles (1996), and the second one was suggested by Liu and Sabatti (2000). Both methodologies will now be introduced.[8]

6.2.2 *Hastings-within-Gibbs step*

In the standard Gibbs sampler, cut-points and latent outcomes are drawn individually from their respective full conditionals, resulting in poor convergence. Instead, Cowles (1996) proposed to sample them jointly through a Hastings-within-Gibbs step. To do so, a set of candidate cut-points $\widetilde{c}$ is drawn from a marginal distribution, and accepted with a given probability:

1. *Sample proposal cut-points.* A normal prior with user-specified precision τ_c is assumed for each cut-point. At each iteration (t) of the Gibbs sampler, each candidate cut-point $\widetilde{c}_l$ is drawn from $\mathcal{N}\left(c_l^{(t-1)}\,;\,\tau_c\right)$ truncated to the interval $\left[\widetilde{c}_{l-1}; c_{l+1}^{(t-1)}\right]$ for $l = 1, ..., L$.

2. *Compute acceptance probability R.* With the set of candidate cut-points in hand, the following acceptance probability is computed:[9]

$$R = \prod_{l=2}^{L-1} \frac{\Phi\left(\tau_c^{1/2}\left[c_{l+1}^{(t-1)} - c_l^{(t-1)}\right]\right) - \Phi\left(\tau_c^{1/2}\left[\widetilde{c}_{l-1} - c_l^{(t-1)}\right]\right)}{\Phi\left(\tau_c^{1/2}\left[\widetilde{c}_{l+1} - \widetilde{c}_l\right]\right) - \Phi\left(\tau_c^{1/2}\left[c_{l-1}^{(t-1)} - \widetilde{c}_l\right]\right)}$$
$$\times \prod_{i=1}^{n} \frac{\Phi(\widetilde{c}_{y_i} - \mu_i) - \Phi(\widetilde{c}_{y_i-1} - \mu_i)}{\Phi\left(c_{y_i}^{(t-1)} - \mu_i\right) - \Phi\left(c_{y_i-1}^{(t-1)} - \mu_i\right)},$$

   where $\mu_i = X_i\beta + \alpha\Theta_i$ is the linear predictor of Y_i^* for individual i, and $\Phi(\cdot)$ represents the cdf of the standard normal distribution.

3. *Accept or reject candidate $\widetilde{c}$.* With probability R, set $c^{(t)} = \widetilde{c}$ and update the latent outcome Y^* accordingly. Otherwise, the cut-points as well as the latent outcome of the previous iteration are kept: $c^{(t)} \equiv c^{(t-1)}$ and $Y^{*(t)} \equiv Y^{*(t-1)}$.

To implement this procedure, a prior precision has to be chosen for the cut-points, thus influencing the acceptance ratio. A rule of thumb consists of choosing τ_c such as the desirable acceptance rate lies between 25% and 50%.

The Hastings step represents a significant improvement of the standard algorithm of Albert and Chib (1993), however convergence might still not be optimal. Furthermore, calculating the acceptance ratio is computationally demanding and slows down the procedure. Because of these potential drawbacks, another method can be considered, namely the group transformation of parameters.

[8] See also Raach (2006); Fahrmeir and Raach (2007).

[9] If the first cut-point is not restricted to zero—when there is no intercept term for instance—, the first product in the expression of R starts at $l = 1$.

6.2.3 Group transformation

This method introduces an intermediate step in the Gibbs sampler which is far less demanding than the Hastings step and demonstrates good convergence performance. It consists of finding an appropriate group transformation of some parameters—possibly all of them—which does not change their target distribution. Hence, the Gibbs sampler remains the same, but its mixing behavior is likely to be greatly improved. Details and theoretical foundations of this methodology are provided in the original paper by Liu and Sabatti (2000), and explained in Raach (2006) and Fahrmeir and Raach (2007). Practically, the standard Gibbs sampler described previously is used to sample the cut-points, and the transformation is then applied to the parameters of interest.

For our purposes, we consider the partial scale group transformation:

$$\Gamma_v = \{\gamma > 0 : \gamma(\omega) = (\gamma\omega_1, ..., \gamma\omega_v, \omega_{v+1}, ..., \omega_W)\},$$

where only the first v elements of the $(W \times 1)$-vector of parameters ω are transformed, the others remain unchanged. Applying the first theorem of Liu and Sabatti (2000), the goal is to sample a suitable parameter γ from a density proportional to $\gamma^{v-1}\pi(\gamma(\omega))$, where $\pi(\cdot)$ represents the density of the parameters to be transformed. In our case, a transformation will be applied on the group:

$$\omega = \{Y_1^*, ..., Y_n^*, \beta_1, ...\beta_b, \alpha_1, ...\alpha_J, c_1, ..., c_{L-1}\},$$

which contains $n + b + J + L - 1$ parameters, where b is the number of covariates in X.

The conditional distribution of this group of parameters is proportional to:

$$\prod_{i=1}^{n}\left[f(Y_i^*|\Theta, X_i, \beta, \alpha)\sum_{l=1}^{L}\mathbb{1}\left[Y_i = l\right]\mathbb{1}\left[c_{l-1} \leqslant Y_i^* < c_l\right]\right] f(\beta, \alpha)f(c), \tag{7}$$

and the target density of the scale parameter can thus be derived as:[10]

$$\begin{aligned}\gamma^{v-1}\pi(\gamma(\omega)) &\propto \gamma^{v-1}\exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(\gamma Y_i^* - X_i\gamma\beta - \gamma\alpha\Theta_i)^2\right\}\exp\left\{-\frac{1}{2}\tau_\alpha(\gamma\alpha)'(\gamma\alpha)\right\},\\ &\propto (\gamma^2)^{\frac{v+1}{2}-1}\exp\left\{-\frac{1}{2}\gamma^2\left[\sum_{i=1}^{n}(Y_i^* - X_i\beta - \alpha\Theta_i)^2 + \tau_\alpha\alpha'\alpha\right]\right\}.\end{aligned} \tag{8}$$

Note that the sum in the expression in brackets in Equation (7) vanishes because it remains constant after the transformation:

$$\sum_{l=1}^{L}\mathbb{1}\left[Y_i = l\right]\mathbb{1}\left[\gamma c_{l-1} \leqslant \gamma Y_i^* < \gamma c_l\right] = \sum_{l=1}^{L}\mathbb{1}\left[Y_i = l\right]\mathbb{1}\left[c_{l-1} \leqslant Y_i^* < c_l\right].$$

[10] To simplify the exposition, noninformative priors are assumed for the slope parameters β and the cut-points c, and therefore their priors are just normalizing constants which disappear from the conditional distribution of γ^2. The factor loadings α are assumed to be a priori normally distributed with mean zero and non-zero precision τ_α, and the kernel of their prior therefore remains. Note that Raach (2006) and Fahrmeir and Raach (2007) implicitly assume noninformative priors also for α since the kernel of its prior distribution vanishes in their formula. The more general case with informative priors would be straightforward to derive.

From Equation (8), it can be seen that γ^2 has to be sampled from a Gamma distribution with parameters:

$$a = \frac{v+1}{2} = \frac{n+b+J+L}{2} \qquad \text{and} \qquad b = \frac{\sum_{i=1}^{n}(Y_i^* - X_i\beta - \alpha\Theta_i)^2 + \tau_\alpha\alpha'\alpha}{2}.$$

The transformation of interest is then carried out by multiplying each element of the group ω by the square root of the sampled γ^2.[11]

In the special case where at least one factor loading is restricted to a constant term, the target density cannot be factorized into Equation (8) as previously, and as a consequence γ^2 cannot be sampled from the same Gamma distribution. We do not consider the case where a factor loading is set to zero as a restriction, because in this case the factor simply vanishes from the distribution and the general case described before can be applied. To solve this problem, the group transformation has to be changed in such a way that the restricted factor loadings are not scaled by γ.

6.3. Mixed error term

Although the use of a mixture of normals for the error term would be straightforward and easy to implement in the framework of our model, there is to the best of our knowledge no prior work on the identification of this type of model. Kottas et al. (2005) have suggested the use of mixtures to extend the traditional ordered response model, introducing more flexibility. But there is no guarantee that this framework would provide reliable results when one or several regressors are latent factors, which can themselves be distributed as mixtures of normals. We therefore leave this for further research, and recommend that the reader be cautious if she wants to deal with a mixed error term in this submodel.

7. CONDITIONAL DISTRIBUTION OF THE FACTORS

Since the latent factors are independently distributed across individuals, the conditional distribution for Θ can be factorized into n conditionals for Θ_1,...,Θ_n. To derive them, note that all the contributions of Θ_i originate from linear regression models:

$$\begin{aligned} Y_{1,i}^* - X_i\beta_1 &= \alpha_1\Theta_i + \varepsilon_{1,i}, \\ &\vdots \\ Y_{S,i}^* - X_i\beta_S &= \alpha_S\Theta_i + \varepsilon_{S,i}. \end{aligned}$$

When the S submodels are stacked, the equation system is of the form:

$$\widetilde{Y}_i^* = \widetilde{\alpha}\Theta_i + \widetilde{\varepsilon}_i, \tag{9}$$

where $\widetilde{Y}_i^*$ is of dimension $(S \times 1)$, $\widetilde{\alpha}$ is a $(S \times J)$-dimensional matrix and Θ is the vector of dimension $(J \times 1)$ containing the factors to be updated. To complete the specification, we have $\widetilde{\varepsilon}_i \sim \mathcal{N}(0\,;\,\Sigma)$. The precision matrix Σ is of dimension $(S \times S)$ and is diagonal since the error terms are independent across equations.[12]

[11] Remember that γ has to be positive in the chosen transformation group Γ.

[12] If the error term is mixed distributed in submodel s, the change of variable explained in Section 2 can be applied, resulting in $\varepsilon_{s,i} \sim \mathcal{N}(0\,;\,1)$.

The conditional distribution for Θ_i is then:

$$f(\Theta_i|\widetilde{Y}_i^*, \widetilde{\alpha}, \Sigma, \psi_\Theta) \propto f(\widetilde{Y}_i^*|\Theta_i, \widetilde{\alpha}, \Sigma) f(\Theta_i|\psi_\Theta), \tag{10}$$

and will depend on the type of distribution assumed for the latent factors—normal or mixed normal.

7.1. Normal case

7.1.1 Multivariate case

If the factors are correlated, they have to be updated simultaneously. Assume they are normally distributed as:

$$\Theta_i \sim \mathcal{N}(0\,;\,\Psi_\Theta)\,,$$

where the precision matrix Ψ_Θ can be considered known since it has been updated in the previous iteration of the Gibbs sampler. The conditional distribution can thus be derived from Equation (10) as:

$$\begin{aligned} f(\Theta_i|\widetilde{Y}_i^*, \widetilde{\alpha}, \Sigma) &\propto \exp\left\{-\frac{1}{2}(\widetilde{Y}_i^* - \widetilde{\alpha}\Theta_i)'\Sigma(\widetilde{Y}_i^* - \widetilde{\alpha}\Theta_i)\right\} \exp\left\{-\frac{1}{2}\Theta_i'\Psi_\Theta\Theta_i\right\}, \\ &\propto \exp\left\{-\frac{1}{2}\left(\Theta_i'\left[\widetilde{\alpha}'\Sigma\widetilde{\alpha} + \Psi_\Theta\right]\Theta_i - 2\Theta_i'\widetilde{\alpha}'\Sigma\widetilde{Y}_i^*\right)\right\}. \end{aligned}$$

Like for the slope parameters of the linear submodel described in Section 2, this last expression provides, after rearranging, the kernel of the following normal distribution:

$$\Theta_i \sim \mathcal{N}\Big(\left[\widetilde{\alpha}'\Sigma\widetilde{\alpha} + \Psi_\Theta\right]^{-1}\widetilde{\alpha}'\Sigma\widetilde{Y}_i^*\,;\,\widetilde{\alpha}'\Sigma\widetilde{\alpha} + \Psi_\Theta\Big)\,.$$

7.1.2 Univariate case

The univariate case is not only of interest when the overall model just has a single latent factor, but also when some of the factors are uncorrelated. The uncorrelated factors can actually be updated one at a time conditional on the other factors. In the system (9), the factors other than the current one being updated just have to be put on the left-hand side of the equation.

A normal distribution with zero mean and precision τ_θ is assumed for the factor, and similarly to the multivariate case, it can be shown that the conditional distribution of θ_i is:

$$\theta_i \sim \mathcal{N}\left(\frac{\widetilde{\alpha}'\Sigma\widetilde{Y}_i^*}{\widetilde{\alpha}'\Sigma\widetilde{\alpha} + \tau_\theta}\,;\,\widetilde{\alpha}'\Sigma\widetilde{\alpha} + \tau_\theta\right).$$

7.2. Mixture of normals case

7.2.1 Multivariate case

The factors are assumed to follow a mixture of K multivariate normal distributions:

$$\Theta_i \sim \sum_{k=1}^{K} p_k \mathcal{N}(\mu_k\,;\,\Psi_k)\,,$$

where μ_k, Ψ_k and p_k represent, respectively, the mean vector, the precision matrix and the weight of the k^{th} mixture component. As in the normal case, their updated values from the previous iteration of the Gibbs sampler are used in the following.

From Equation (10), the conditional is derived as:

$$
\begin{aligned}
&f(\Theta_i|\widetilde{Y}_i^*,\widetilde{\alpha},\Sigma) \propto f(\widetilde{Y}_i^*|\Theta_i,\widetilde{\alpha},\Sigma)f(\Theta_i),\\
&\propto \exp\left\{-\frac{1}{2}(\widetilde{Y}_i^*-\widetilde{\alpha}\Theta_i)'\Sigma(\widetilde{Y}_i^*-\widetilde{\alpha}\Theta_i)\right\}\sum_{k=1}^{K}p_k|\Psi_k|^{1/2}\exp\left\{-\frac{1}{2}(\Theta_i-\mu_k)'\Psi_k(\Theta_i-\mu_k)\right\},\\
&\propto \sum_{k=1}^{K}p_k|\Psi_k|^{1/2}\exp\left\{-\frac{1}{2}\left(\Theta_i'\widetilde{\alpha}'\Sigma\widetilde{\alpha}\Theta_i-2\Theta_i'\widetilde{\alpha}'\Sigma\widetilde{Y}_i^*+\Theta_i'\Psi_k\Theta_i-2\Theta_i'\Psi_k\mu_k+\mu_k'\Psi_k\mu_k\right)\right\},\\
&\propto \sum_{k=1}^{K}p_k|\Psi_k|^{1/2}\exp\left\{-\frac{1}{2}\left(\Theta_i'\left[\widetilde{\alpha}'\Sigma\widetilde{\alpha}+\Psi_k\right]\Theta_i-2\Theta_i'\left[\widetilde{\alpha}'\Sigma\widetilde{Y}_i^*+\Psi_k\mu_k\right]+\mu_k'\Psi_k\mu_k\right)\right\},\\
&\propto \sum_{k=1}^{K}p_k|\Psi_k|^{1/2}\exp\left\{-\frac{1}{2}\left(\mu_k'\Psi_k\mu_k-\left[\widetilde{\alpha}'\Sigma\widetilde{Y}_i^*+\Psi_k\mu_k\right]'\left[\widetilde{\alpha}'\Sigma\widetilde{\alpha}+\Psi_k\right]^{-1}[\bullet]\right)\right\}\\
&\qquad\times\exp\left\{-\frac{1}{2}\left(\Theta_i-\left[\widetilde{\alpha}'\Sigma\widetilde{\alpha}+\Psi_k\right]^{-1}\left[\widetilde{\alpha}'\Sigma\widetilde{Y}_i^*+\Psi_k\mu_k\right]\right)'\left(\widetilde{\alpha}'\Sigma\widetilde{\alpha}+\Psi_k\right)(\bullet)\right\},
\end{aligned}
\tag{11}
$$

where $(\bullet)$ represents the first factor of the corresponding sandwich matrix. The kernel of a mixture of multivariate normal distributions emerges from Equation (11). To further simplify its expression, the first exponential can be factorized with respect to μ_k, and after some matrix algebra, it can be shown to be proportional to:[13]

$$
\exp\left\{-\frac{1}{2}\left(\mu_k'\Psi_k\mu_k-\left[\widetilde{\alpha}'\Sigma\widetilde{Y}_i^*+\Psi_k\mu_k\right]'\left[\widetilde{\alpha}'\Sigma\widetilde{\alpha}+\Psi_k\right]^{-1}[\bullet]\right)\right\} \tag{12}
$$

$$
\propto \exp\left\{-\frac{1}{2}\left(\mu_k-\left[\widetilde{\alpha}'\Sigma\widetilde{\alpha}\right]^{-1}\widetilde{\alpha}'\Sigma\widetilde{Y}_i^*\right)'\left(\Psi_k\left[\widetilde{\alpha}'\Sigma\widetilde{\alpha}+\Psi_k\right]^{-1}\widetilde{\alpha}'\Sigma\widetilde{\alpha}\right)(\bullet)\right\}. \tag{13}
$$

Using this result and writing $A \equiv \widetilde{\alpha}'\Sigma\widetilde{\alpha}$ and $B \equiv \widetilde{\alpha}'\Sigma\widetilde{Y}_i^*$ to make the notation clearer, Equation (11) can be re-expressed as:

$$
\begin{aligned}
&\sum_{k=1}^{K}p_k|\Psi_k|^{1/2}\exp\left\{-\frac{1}{2}\left(\mu_k-A^{-1}B\right)'\left(\Psi_k\left(A+\Psi_k\right)^{-1}A\right)(\bullet)\right\}\\
&\qquad\qquad\times\exp\left\{-\frac{1}{2}\left(\Theta_i-[A+\Psi_k]^{-1}[B+\Psi_k\mu_k]\right)'(A+\Psi_k)(\bullet)\right\},\\
&\propto \sum_{k=1}^{K}p_k\left|\Psi_k(A+\Psi_k)^{-1}A\right|^{1/2}\exp\left\{-\frac{1}{2}\left(\mu_k-A^{-1}B\right)'\left(\Psi_k\left[A+\Psi_k\right]^{-1}A\right)(\bullet)\right\}\\
&\qquad\qquad\times|A+\Psi_k|^{1/2}\exp\left\{-\frac{1}{2}\left(\Theta_i-[A+\Psi_k]^{-1}[B+\Psi_k\mu_k]\right)'(A+\Psi_k)(\bullet)\right\}.
\end{aligned}
$$

Hence, the conditional of the factors Θ_i is a mixture of multivariate normal distributions with the following means, precisions and weights:

$$
\mu_k^{post} = \left[\widetilde{\alpha}'\Sigma\widetilde{\alpha}+\Psi_k\right]^{-1}\left[\widetilde{\alpha}'\Sigma\widetilde{Y}_i^*+\Psi_k\mu_k\right],
$$

[13] See Appendix A for details.

$$\Psi_k^{post} = \widetilde{\alpha}'\Sigma\widetilde{\alpha} + \Psi_k,$$
$$p_k^{post} \propto p_k \; \phi\Big(\mu_k \,;\, \left[\widetilde{\alpha}'\Sigma\widetilde{\alpha}\right]^{-1} \widetilde{\alpha}'\Sigma\widetilde{Y}_i^* \,,\, \Psi_k \left[\widetilde{\alpha}'\Sigma\widetilde{\alpha} + \Psi_k\right]^{-1} \widetilde{\alpha}'\Sigma\widetilde{\alpha}\Big),$$

where $\phi(\cdot)$ is the pdf of the multivariate normal distribution with given mean and precision evaluated at μ_k.

7.2.2 Univariate case

As in the normal case, the univariate result follows from the multivariate derivation in the previous section. Assuming that the factor follows a mixture of K univariate normals:

$$\theta_i \sim \sum_{k=1}^{K} p_k \mathcal{N}(\mu_k \,;\, \tau_k)\,,$$

where μ_k, τ_k and p_k are, respectively, the mean, the precision and the weight of mixture component k, the conditional distribution can be shown to be a mixture of univariate normal distributions with the following parameters:

$$\mu_k^{post} = \frac{\widetilde{\alpha}'\Sigma\widetilde{Y}_i^* + \tau_k\mu_k}{\widetilde{\alpha}'\Sigma\widetilde{\alpha} + \tau_k},$$
$$\tau_k^{post} = \widetilde{\alpha}'\Sigma\widetilde{\alpha} + \tau_k,$$
$$p_k^{post} \propto p_k \; \phi\left(\mu_k \,;\, \frac{\widetilde{\alpha}'\Sigma\widetilde{Y}_i^*}{\widetilde{\alpha}'\Sigma\widetilde{\alpha}} \,,\, \frac{\tau_k\widetilde{\alpha}'\Sigma\widetilde{\alpha}}{\widetilde{\alpha}'\Sigma\widetilde{\alpha} + \tau_k}\right).$$

8. UPDATING FACTOR DISTRIBUTION PARAMETERS

Once the latent factors have been sampled, they can be regarded as known and the parameters of their distribution can be updated.

8.1. Normal case

The latent factors are assumed to be centered at zero and thus only their precisions have to be updated.

8.1.1 Univariate case

When $\theta \sim \mathcal{N}(0\,;\,\tau_\theta)$, the precision τ_θ is sampled exactly in the same way as the precision of the error term in the linear submodel (see Section 2). A conjugate Gamma distribution $\mathcal{G}a(g_1\,;\,g_2)$ is assumed, and the conditional of the precision can be shown to be:

$$\tau_\theta \sim \mathcal{G}a\left(g_1 + \frac{n}{2}\,;\, g_2 + \frac{1}{2}\sum_i \theta_i^2\right).$$

8.1.2 Multivariate case

In the multivariate case, we specify $\Theta \sim \mathcal{N}(0\,;\,\Psi_\Theta)$. The precision matrix is assumed to be a priori Wishart distributed with v degrees of freedom and scale matrix Γ:

$$\Psi_\Theta \sim \mathcal{W}(v\,;\,\Gamma),$$

and applying Bayes' rule, the conditional distribution is derived as:[14]

$$\begin{aligned}
&f(\Psi_\Theta|\Theta) \propto f(\Theta|\Psi_\Theta) f(\Psi_\Theta),\\
&\propto |\Psi_\Theta|^{n/2} \exp\left\{-\frac{1}{2}\sum_i \Theta_i'\Psi_\Theta\Theta_i\right\} |\Psi_\Theta|^{(v-J-1)/2} \exp\left\{-\frac{1}{2}\mathrm{tr}\left(\Gamma^{-1}\Psi_\Theta\right)\right\},\\
&\propto |\Psi_\Theta|^{(n+v-J-1)/2} \exp\left\{-\frac{1}{2}\left[\sum_i \Theta_i'\Psi_\Theta\Theta_i + \mathrm{tr}\left(\Gamma^{-1}\Psi_\Theta\right)\right]\right\},\\
&\propto |\Psi_\Theta|^{(n+v-J-1)/2} \exp\left\{-\frac{1}{2}\left[\mathrm{tr}\left(\left[\sum_i \Theta_i\Theta_i' + \Gamma^{-1}\right]\Psi_\Theta\right)\right]\right\}.
\end{aligned}$$

Hence, the conditional of the precision matrix is a Wishart distribution:

$$\Psi_\Theta \sim \mathcal{W}\left(n+v\,;\,\left[\sum_i \Theta_i\Theta_i' + \Gamma^{-1}\right]^{-1}\right).$$

8.2. Mixture of normals case

When the distribution of the latent factors is specified to be a mixture of normals, the sampling procedure is slightly more complicated since different sets of parameters—group indicators, components' means, precisions and weights—have to be sequentially updated. The sampling scheme is the same as the one used to update the parameters of the error term distribution of the submodels in the mixed case, and is presented in Section 9.

9. UPDATING THE PARAMETERS OF A NORMAL MIXTURE DISTRIBUTION

This section describes how to update the parameters of a normal mixture distribution under zero-mean restriction. This procedure is used to sample the parameters of the error term in the mixed case, but also to sample the parameters of the factor distribution in the mixed case.

9.1. Multivariate case

Let Z_i be a random vector of dimension $(J \times 1)$ distributed as a mixture of K multivariate normal distributions:

$$Z_i \sim \sum_{k=1}^{K} p_k \mathcal{N}(\mu_k\,;\,\Psi_k)\,,$$

[14] The result is obtained using a basic property of the trace operator: $\mathrm{tr}\,(ABC) = \mathrm{tr}\,(CAB)$, where A, B and C are matrices of appropriate sizes.

where μ_k is the vector of means of dimension $(J \times 1)$, Ψ_k is the $(J \times J)$-dimensional precision matrix, and p_k is the weight of mixture component k.

We adopt the group-indicator approach (Diebolt and Robert 1994), and first update the group membership of each individual i. Once we know to which group each individual belongs, the different sets of mixture parameters are sampled component-wise. For this purpose, let $g_{k,i}$ be the group indicator which is equal to one if individual i belongs to group k, and to zero otherwise.

9.1.1 Updating group indicators

Since each individual can belong to only one mixture group, the probability mass of each $g_{k,i}$ is simply the corresponding weight p_k. As a consequence, the vector of indicators $g_i = (g_{1,i}, ..., g_{K,i})'$ follows a multinomial distribution with the mixture weights as parameters. Conditional on group membership, Z_i is normally distributed, and the posterior of the group indicators is, for individual i and for all $k = 1, ..., K$:

$$\begin{aligned} f(g_{k,i} = 1|Z_i) &\propto f(Z_i|g_{k,i} = 1)f(g_{k,i} = 1), \\ &\propto \phi(Z_i; \mu_k, \Psi_k)p_k, \end{aligned}$$

where $\phi(Z_i; \mu_k, \Psi_k)$ denotes the pdf of the multivariate normal distribution with mean μ_k and precision matrix Ψ_k evaluated at Z_i.

Thus, group membership probabilities are computed as follows in order to ensure normalization:

$$Pr(g_{k,i} = 1) = \frac{p_k \phi(Z_i; \mu_k, \Psi_k)}{\sum_{l=1}^{K} p_l \phi(Z_i; \mu_l, \Psi_l)}, \tag{14}$$

and the conditional of group indicators is therefore the following multinomial distribution:

$$g_i \sim \mathcal{M}\left(1\,;\, \frac{p_1 \phi(Z_i; \mu_1, \Psi_1)}{\sum_{l=1}^{K} p_l \phi(Z_i; \mu_l, \Psi_l)}, ..., \frac{p_K \phi(Z_i; \mu_K, \Psi_K)}{\sum_{l=1}^{K} p_l \phi(Z_i; \mu_l, \Psi_l)}\right).$$

9.1.2 Updating mixture precisions

Once the group membership of each individual has been sampled, the precision matrices of the mixture components can be updated exactly in the same manner as in the normal case described in Section 8. Assuming a Wishart prior distribution with v degrees of freedom and with scale matrix Γ for mixture precision Ψ_k:

$$\Psi_k \sim \mathcal{W}(v\,;\, \Gamma),$$

the conditional can be shown to be the following Wishart distribution:

$$\Psi_k \sim \mathcal{W}\left(v + n_k\,;\, \left[\sum_{i:g_{k,i}=1} (Z_i - \mu_k)(Z_i - \mu_k)' + \Gamma^{-1}\right]^{-1}\right),$$

where the sum in the scale parameter is over the individuals belonging to mixture group k, and n_k stands for the number of individuals in this group.

9.1.3 Updating mixture weights

Let $p = (p_1, ..., p_K)'$ and $g = (g_1, ..., g_n)'$. A Dirichlet distribution is assumed for the prior of the weights:

$$p \sim \mathcal{D}(a, ..., a)\,.$$

Conditional on g_i, Z_i appears to be normally distributed in each mixture group and we can write $f(Z_i|g_i, p) = \prod_{k=1}^{K}[p_k\phi(Z_i; \mu_k, \Psi_k)]^{g_{k,i}}$. With this expression in hand, the conditional of the weights can be derived as:

$$\begin{aligned} f(p|Z, g) &\propto f(Z|g, p)f(p|g), \\ &\propto \prod_{i=1}^{n}\prod_{k=1}^{K} p_k^{g_{k,i}} \prod_{k=1}^{K} p_k^{a-1}, \\ &\propto \prod_{k=1}^{K} p_k^{n_k+a-1}, \end{aligned}$$

providing the kernel of a Dirichlet distribution. Hence, $p \sim \mathcal{D}(n_1 + a, ..., n_K + a)$.

9.1.4 Updating mixture means

Each mean vector is assumed to be a priori normally distributed:

$$\mu_k \sim \mathcal{N}\big(\mu_k^0\,;\, \Omega_k^0\big)\,,$$

where μ_k and μ_k^0 are of dimension $(J \times 1)$, and Ω_k^0 is a $(J \times J)$-dimensional matrix.

The mean of the mixture is restricted to zero as follows:

$$\sum_{k=1}^{K} p_k\mu_k = 0 \qquad \Leftrightarrow \qquad \mu_K = -\sum_{k=1}^{K-1} \frac{p_k}{p_K}\mu_k. \tag{15}$$

Because of this restriction, only $K - 1$ mixture means actually have to be updated, and the last one can then be computed given the other component means and the mixture weights. The idea is thus to first update the first $K - 1$ mean vectors using Bayes' rule, imposing the restriction on both the likelihood and prior of the means, and finally computing the last mean vector so as to fulfill the zero-mean constraint. Two different approaches can be implemented: the mean vectors can be either drawn sequentially across mixture components, or simultaneously.

Drawing the means sequentially. Mean vectors are drawn component-wise, holding constant the means other than the one being updated. To do so, the conditional of the l^{th} mean vector has to be derived conditional on the means of the other components. Let μ_{-l} be the set of mixture mean vectors excluding the l^{th} one. Then, replacing μ_K with its expression given by Equation (15), the conditional distribution is, for $l = 1, ..., K - 1$:

$$f(\mu_l|Z, \mu_{-l}, \Psi, p) \propto f(Z|\mu_l, \mu_{-l}, \Psi, p)f(\mu_l, \mu_{-l}),$$

$$
\begin{aligned}
&\propto \prod_{k=1}^{K-1} \exp\left\{-\frac{1}{2}\sum_{i:g_{k,i}=1}(Z_i-\mu_k)'\Psi_k(Z_i-\mu_k)\right\}\exp\left\{-\frac{1}{2}(\mu_k-\mu_k^0)'\Omega_k^0(\mu_k-\mu_k^0)\right\}\\
&\quad\times\exp\left\{-\frac{1}{2}\sum_{i:g_{K,i}=1}\left(Z_i+\sum_{k=1}^{K-1}\frac{p_k}{p_K}\mu_k\right)'\Psi_K\left(Z_i+\sum_{k=1}^{K-1}\frac{p_k}{p_K}\mu_k\right)\right\}\\
&\quad\times\exp\left\{-\frac{1}{2}\left(-\sum_{k=1}^{K-1}\frac{p_k}{p_K}\mu_k-\mu_K^0\right)'\Omega_K^0\left(-\sum_{k=1}^{K-1}\frac{p_k}{p_K}\mu_k-\mu_K^0\right)\right\},\\
&\propto \exp\left\{-\frac{1}{2}\sum_{i:g_{l,i}=1}(Z_i-\mu_l)'\Psi_l(Z_i-\mu_l)\right\}\exp\left\{-\frac{1}{2}(\mu_l-\mu_l^0)'\Omega_l^0(\mu_l-\mu_l^0)\right\}\\
&\quad\times\exp\left\{-\frac{1}{2}\left(\frac{p_l}{p_K}\right)^2\sum_{i:g_{K,i}=1}\left(\mu_l+\left[\sum_{\substack{k=1\\k\neq l}}^{K-1}\frac{p_k}{p_l}\mu_k+\frac{p_K}{p_l}Z_i\right]\right)'\Psi_K(\bullet)\right\}\\
&\quad\times\exp\left\{-\frac{1}{2}\left(\frac{p_l}{p_K}\right)^2\left(\mu_l+\left[\sum_{\substack{k=1\\k\neq l}}^{K-1}\frac{p_k}{p_l}\mu_k+\frac{p_K}{p_l}\mu_K^0\right]\right)'\Omega_K^0(\bullet)\right\},\\
&\propto \exp\left\{-\frac{1}{2}\left(n_l\mu_l'\Psi_l\mu_l-2\mu_l'\Psi_l\sum_{i:g_{l,i}=1}Z_i\right)\right\}\exp\left\{-\frac{1}{2}\left(\mu_l'\Omega_l^0\mu_l-2\mu_l'\Omega_l^0\mu_l^0\right)\right\}\\
&\quad\times\exp\left\{-\frac{1}{2}\left(\frac{p_l}{p_K}\right)^2\left[n_K\mu_l'\Psi_K\mu_l+2\mu_l'\Psi_K\left(n_K\sum_{\substack{k=1\\k\neq l}}^{K-1}\frac{p_k}{p_l}\mu_k+\frac{p_K}{p_l}\sum_{i:g_{K,i}=1}Z_i\right)\right]\right\}\\
&\quad\times\exp\left\{-\frac{1}{2}\left(\frac{p_l}{p_K}\right)^2\left[\mu_l'\Omega_K^0\mu_l+2\mu_l'\Omega_K^0\left(\sum_{\substack{k=1\\k\neq l}}^{K-1}\frac{p_k}{p_l}\mu_k+\frac{p_K}{p_l}\mu_K^0\right)\right]\right\},\\
&\propto \exp\left\{-\frac{1}{2}\left(\mu_l'\left[n_l\Psi_l+\Omega_l^0+\left(\frac{p_l}{p_K}\right)^2\left(n_K\Psi_K+\Omega_K^0\right)\right]\mu_l\right.\right.\\
&\quad-2\mu_l'\left[\Psi_l\sum_{i:g_{l,i}=1}Z_i+\Omega_l^0\mu_l^0-\left(\frac{p_l}{p_K}\right)^2\left(n_K\Psi_K+\Omega_K^0\right)\sum_{\substack{k=1\\k\neq l}}^{K-1}\frac{p_k}{p_l}\mu_k\right.\\
&\quad\left.\left.\left.-\frac{p_l}{p_K}\left(\Psi_K\sum_{i:g_{K,i}=1}Z_i+\Omega_K^0\mu_K^0\right)\right]\right)\right\},
\end{aligned}
$$

where n_l $(l=1,...,K)$ is the number of observations in mixture group l.

This last expression can be factorized to provide the kernel of a multivariate normal

distribution with the following precision and mean, for $l = 1, ..., K-1$:

$$\Omega_{\mu_l}^{post} = n_l \Psi_l + \Omega_l^0 + \left(\frac{p_l}{p_K}\right)^2 \left(n_K \Psi_K + \Omega_K^0\right),$$

$$\mu_{\mu_l}^{post} = \left[\Omega_{\mu_l}^{post}\right]^{-1} \left[\Psi_l \sum_{i:g_{l,i}=1} Z_i + \Omega_l^0 \mu_l^0 - \left(\frac{p_l}{p_K}\right)^2 \left(n_K \Psi_K + \Omega_K^0\right) \sum_{\substack{k=1 \\ k \neq l}}^{K-1} \frac{p_k}{p_l} \mu_k \right.$$
$$\left. - \frac{p_l}{p_K} \left(\Psi_K \sum_{i:g_{K,i}=1} Z_i + \Omega_K^0 \mu_K^0 \right) \right].$$

Once the first $K-1$ mean vectors have been sampled, the last mean μ_K is computed to fulfill the zero-mean restriction of the mixture given by Equation (15).

Drawing the means simultaneously. To make the calculations easier to follow, compact form matrices will be used henceforth. For this purpose, some notational conventions have to be introduced before going any further. We define the following stacked vectors and block-matrices:

$$\begin{aligned}
&p_{-K} = (p_1, ..., p_{K-1})', && \widetilde{p} = p_{-K} \otimes I_J, \\
&\widetilde{\mu} = (\mu_1', ..., \mu_{K-1}')', && \widetilde{\mu}^0 = (\mu_1^{0\prime}, ..., \mu_{K-1}^{0}{}')', \\
&\widetilde{\Psi} = \operatorname{diag}_{k=1}^{K-1}(\Psi_k), && \widetilde{\Omega}^0 = \operatorname{diag}_{k=1}^{K-1}\left(\Omega_k^0\right), \\
&\widetilde{g}_i = (g_{1,i}, ..., g_{K-1,i})', && \widetilde{Z}_i = \iota_{K-1} \otimes Z_i,
\end{aligned}$$

where $\otimes$ denotes the Kronecker product, I_J is the identity matrix of dimension $(J \times J)$, ι_{K-1} is the $(K-1)$-dimensional vector of 1's, and diag $(\cdot)$ is the matrix operator stacking the specified elements to create a block-diagonal matrix.

Using this notation, the zero-mean restriction of Equation (15) can be rewritten as:

$$\mu_K = -\frac{1}{p_K} \widetilde{p}' \widetilde{\mu}. \tag{16}$$

Since the conditional distribution will be derived given group membership, the following selection matrix will be useful for picking up the parameters corresponding to the mixture group individual i belongs to:

$$G_i = (\widetilde{g}_i \, \widetilde{g}_i{}') \otimes I_J.$$

Replacing the last mean vector μ_K with its expression from Equation (16) in the likelihood and in the prior, the conditional distribution of the first $K-1$ means is:

$$f(\widetilde{\mu}|Z, p, \Psi) \propto f(Z|\widetilde{\mu}, p, \Psi) f(\widetilde{\mu}),$$
$$\propto \exp\left\{ -\frac{1}{2} \sum_{i=1}^{n} \left[(\widetilde{Z}_i - \widetilde{\mu})' G_i \widetilde{\Psi} (\widetilde{Z}_i - \widetilde{\mu}) + g_{K,i} \left(Z_i + \frac{1}{p_K} \widetilde{p}' \widetilde{\mu} \right)' \Psi_K \left(Z_i + \frac{1}{p_K} \widetilde{p}' \widetilde{\mu} \right) \right] \right\}$$
$$\times \exp\left\{ -\frac{1}{2} \left[(\widetilde{\mu} - \widetilde{\mu}^0)' \widetilde{\Omega}^0 (\widetilde{\mu} - \widetilde{\mu}^0) + \left(-\frac{1}{p_K} \widetilde{p}' \widetilde{\mu} - \mu_K^0 \right)' \Omega_K^0 \left(-\frac{1}{p_K} \widetilde{p}' \widetilde{\mu} - \mu_K^0 \right) \right] \right\},$$

$$
\begin{aligned}
&\propto \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\left[\widetilde{\mu}' G_i \widetilde{\Psi}\widetilde{\mu} - 2\widetilde{\mu}' G_i \widetilde{\Psi}\widetilde{Z}_i + \frac{g_{K,i}}{p_K^2}\widetilde{\mu}'\widetilde{p}\,\Psi_K \widetilde{p}'\widetilde{\mu} + 2\frac{g_{K,i}}{p_K}\widetilde{\mu}'\widetilde{p}\,\Psi_K Z_i\right]\right\} \\
&\quad \times \exp\left\{-\frac{1}{2}\left[\widetilde{\mu}'\widetilde{\Omega}^0\widetilde{\mu} - 2\widetilde{\mu}'\widetilde{\Omega}^0\widetilde{\mu}^0 + \frac{1}{p_K^2}\widetilde{\mu}'\widetilde{p}\,\Omega_K^0\widetilde{p}'\widetilde{\mu} + 2\frac{1}{p_K}\widetilde{\mu}'\widetilde{p}\,\Omega_K^0\mu_K^0\right]\right\}, \\
&\propto \exp\left\{-\frac{1}{2}\left[\widetilde{\mu}'\left(\sum_{i=1}^{n}\left[G_i\widetilde{\Psi} + \frac{g_{K,i}}{p_K^2}\widetilde{p}\,\Psi_K\widetilde{p}'\right] + \widetilde{\Omega}^0 + \frac{1}{p_K^2}\widetilde{p}\,\Omega_K^0\widetilde{p}'\right)\widetilde{\mu}\right.\right. \\
&\quad \left.\left. -2\widetilde{\mu}'\left(\sum_{i=1}^{n}\left[G_i\widetilde{\Psi}\widetilde{Z}_i - \frac{g_{K,i}}{p_K}\widetilde{p}\,\Psi_K Z_i\right] + \widetilde{\Omega}^0\widetilde{\mu}^0 - \frac{1}{p_K}\widetilde{p}\,\Omega_K^0\mu_K^0\right)\right]\right\}.
\end{aligned}
$$

This last expression can be factorized into the kernel of a multivariate normal distribution. Its precision matrix is:

$$
\begin{aligned}
\widetilde{\Omega}^{post} &= \sum_{i=1}^{n}\left[G_i\widetilde{\Psi} + \frac{g_{K,i}}{p_K^2}\widetilde{p}\,\Psi_K\widetilde{p}'\right] + \widetilde{\Omega}^0 + \frac{1}{p_K^2}\widetilde{p}\,\Omega_K^0\widetilde{p}', \\
&= \mathrm{diag}_{k=1}^{K-1}\left(n_k\Psi_k + \Omega_k^0\right) + \frac{1}{p_K^2}\widetilde{p}\,(n_K\Psi_K + \Omega_K^0)\widetilde{p}',
\end{aligned}
$$

where n_k $(k = 1, ..., K)$ is the number of observations in mixture group k. Hence, the precision of the conditional is a block matrix defined as follows, for all $k, l = 1, ..., K-1$ and $k \neq l$:

$$
\begin{aligned}
\widetilde{\Omega}^{post}_{[k,k]} &= n_k\Psi_k + \Omega_k^0 + \left(\frac{p_k}{p_K}\right)^2(n_K\Psi_K + \Omega_K^0) && \text{for diagonal block elements,} \\
\widetilde{\Omega}^{post}_{[k,l]} &= \frac{p_k p_l}{p_K^2}(n_K\Psi_K + \Omega_K^0) && \text{for off-diagonal block elements.}
\end{aligned}
$$

As for the mean of the conditional, it is equal to:

$$
\begin{aligned}
\widetilde{\mu}^{post} &= \left[\widetilde{\Omega}^{post}\right]^{-1}\left(\sum_{i=1}^{n}\left[G_i\widetilde{\Psi}\widetilde{Z}_i - \frac{g_{K,i}}{p_K}\widetilde{p}\,\Psi_K Z_i\right] + \widetilde{\Omega}^0\widetilde{\mu}^0 - \frac{1}{p_K}\widetilde{p}\,\Omega_K^0\mu_K^0\right), \\
&= \left[\widetilde{\Omega}^{post}\right]^{-1}\mathrm{vec}_{k=1}^{K-1}\left(\Psi_k\sum_{i:g_{k,i}=1}Z_i + \Omega_k^0\mu_k^0 - \frac{p_k}{p_K}\left[\Psi_K\sum_{i:g_{K,i}=1}Z_i + \Omega_K^0\mu_K^0\right]\right),
\end{aligned}
$$

where vec $(\cdot)$ is the operator vertically stacking the specified elements to create a vector.

As in the previous case, the last mean vector μ_K is computed to fulfill the zero-mean restriction of the mixture given by Equation (15).

9.2. Univariate case

The following results are directly derived from the multivariate case. The same notation is used here. Assume that z_i is univariate and mixed distributed:

$$
z_i \sim \sum_{k=1}^{K} p_k\mathcal{N}(\mu_k\,;\,\tau_k)\,,
$$

where μ_k and τ_k are both scalars representing the mean and the precision of mixture component k, and p_k is the corresponding weight.

As in the multivariate case, group indicators are drawn from a multinomial distribution:

$$g_i \sim \mathcal{M}\left(1\,;\, \frac{p_1\phi(z_i;\mu_1,\tau_1)}{\sum_{l=1}^{K} p_l\phi(z_i;\mu_l,\tau_l)},...,\frac{p_K\phi(z_i;\mu_K,\tau_K)}{\sum_{l=1}^{K} p_l\phi(z_i;\mu_l,\tau_l)}\right).$$

Given that z_i is normally distributed in each mixture group, a conjugate Gamma prior $\mathcal{G}a(g_1\,;\,g_2)$ can be assumed for the precision τ_k, and like in the normal case (Section 8) the conditional is derived as:

$$\tau_k \sim \mathcal{G}a\left(g_1+\frac{n_k}{2}\,;\, g_2+\frac{1}{2}\sum_{i:g_{k,i}=1} z_i^2\right).$$

For the means, the same procedure as in the multivariate case is implemented to achieve the zero-mean restriction. Assuming that each mean is a priori normally distributed ($\mu_k \sim \mathcal{N}\left(\mu_k^0\,;\,\omega_k^0\right)$, $k=1,...K$), the first $K-1$ means are simultaneously drawn from their conditional, which is a multivariate normal distribution with precision and mean:

$$\widetilde{\tau}^{post} = \operatorname{diag}_{k=1}^{K-1}\left(n_k\tau_k+\omega_k^0\right)+\frac{n_K\tau_K+\omega_K^0}{p_K^2}p_{-K}p_{-K}',$$

$$\widetilde{\mu}^{post} = \left[\widetilde{\tau}^{post}\right]^{-1}\operatorname{vec}_{k=1}^{K-1}\left(\tau_k\sum_{i:g_{k,i}=1} z_i+\omega_k^0\mu_k^0-\frac{p_k}{p_K}\left[\tau_K\sum_{i:g_{K,i}=1} z_i+\omega_K^0\mu_K^0\right]\right),$$

and the last mean μ_K is finally computed according to Equation (15).

10. CONCLUSION

The purpose of this article was to provide all the technical details required to construct the Gibbs sampler for the estimation of factor structure models. Within the proposed framework, a wide range of submodels including linear, dichotomous and censored response submodels, as well as unordered response and ordered response submodels can be easily accommodated. We derived all the conditional distributions of interest. Having the conditionals in hand, it is straightforward to construct the Gibbs sampler step by step.

ACKNOWLEDGMENTS

I am very grateful to James Heckman and to the members of his research group at the Economics Research Center, University of Chicago, for stimulating and enlightening discussions. I am also very thankful to François Laisney for helpful comments. Financial support by the Baden-Württemberg (LGFG scholarship) and by the German Academic Exchange Service (DAAD) is gratefully acknowledged.

References

Albert, J. H. and S. Chib (1993): "Bayesian Analysis of Binary and Polychotomous Response Data," *Journal of the American Statistical Association*, 88, 669–679.

Bunch, D. S. (1991): "Estimability in the multinomial probit model," *Transportation Research Part B: Methodological*, 25, 1–12.

Carneiro, P., K. T. Hansen, and J. J. Heckman (2003): "Estimating Distributions of Treatment Effects with an Application to the Returns to Schooling and Measurement of the Effects of Uncertainty on College," Working Paper No. 9546, National Bureau of Economic Research.

Casella, G. and E. I. George (1992): "Explaining the Gibbs Sampler," *The American Statistician*, 46, 167–174.

Chib, S. (1992): "Bayes inference in the Tobit censored regression model," *Journal of Econometrics*, 51, 79–99.

Cowles, M. K. (1996): "Accelerating Monte Carlo Markov Chain Convergence for Cumulative-link Generalized Linear Models," *Statistics and Computing*, 6, 101–111.

Diebolt, J. and C. P. Robert (1994): "Estimation of Finite Mixture Distributions Through Bayesian Sampling," *Journal of the Royal Statistical Society Series B: Methodological*, 56, 363–375.

Fahrmeir, L. and A. W. Raach (2007): "A Bayesian Semiparametric Latent Variable Model for Mixed Responses," *Psychometrika*, 72, 327–346.

Geweke, J. F. and M. P. Keane (1999): "Mixture of Normals Probit Models," in *Analysis of panels and limited dependent variables: a volume in honor of G. S. Maddala*, ed. by C. Hsiao, G. S. Maddala, K. Lahiri, and L. F. Lee, Cambridge University Press, 49–78.

Hansen, K. T., J. J. Heckman, and K. J. Mullen (2004): "The Effect of Schooling and Ability on Achievement Test Scores," *Journal of Econometrics*, 121, 39–98.

Heckman, J. J., J. Stixrud, and S. Urzua (2006): "The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior," *Journal of Labor Economics*, 24, 411–482.

Keane, M. P. (1992): "A Note on Identification in the Multinomial Probit Model," *Journal of Business & Economic Statistics*, 10, 193–200.

Kottas, A., P. Müller, and F. Quintana (2005): "Nonparametric Bayesian Modeling for Multivariate Ordinal Data," *Journal of Computational and Graphical Statistics*, 14, 610–625.

Liu, J. L. and C. Sabatti (2000): "Generalised Gibbs Sampler and Multigrid Monte Carlo for Bayesian Computation," *Biometrika*, 87, 353–369.

McCulloch, R. E. and P. E. Rossi (2000): "Bayesian analysis of the multinomial probit model," in *Simulation-based inference in econometrics: methods and applications*, ed. by R. S. Mariano, T. Schuermann, and M. J. Weeks, Cambridge University Press, 158–178.

Raach, A. W. (2006): "A Bayesian semiparametric latent variable model for binary, ordinal and continuous response," PhD thesis, Ludwig-Maximilians-Universität München.

Tanner, M. A. and W. H. Wong (1987): "The Calculation of Posterior Distributions by Data Augmentation," *Journal of the American Statistical Association*, 82, 528–540.

van Dyk, D. A. and X.-L. Meng (2001): "The Art of Data Augmentation," *Journal of Computational & Graphical Statistics*, 10, 1–50.

APPENDIX A: Some matrix algebra

Let E and F be two invertible square matrices of same dimension. The following results hold:

$$\begin{aligned} E - E(F+E)^{-1}E &= E(F+E)^{-1}(F+E) - E(F+E)^{-1}E, \\ &= E(F+E)^{-1}(F+E-E), \\ &= E(F+E)^{-1}F. \end{aligned}$$

Given that $(EF)^{-1} = F^{-1}E^{-1}$, it can be shown that:

$$\begin{aligned} &(E+F)^{-1}F[F(E+F)^{-1}E]^{-1}F(E+F)^{-1} + (E+F)^{-1} \\ &\quad = (E+F)^{-1}FE^{-1} + (E+F)^{-1}, \\ &\quad = (E+F)^{-1}(FE^{-1} + I), \\ &\quad = (E+F)^{-1}(F+E)E^{-1}, \\ &\quad = E^{-1}. \end{aligned}$$

Using these identities, Equation (12) can be simplified as follows:

$$\begin{aligned} &\exp\left\{-\frac{1}{2}\left[\mu_k'\Psi_k\mu_k - (B+\Psi_k\mu_k)'\left(A+\Psi_k\right)^{-1}(\bullet)\right]\right\} \\ &= \exp\left\{-\frac{1}{2}\left[\mu_k'\left(\Psi_k - \Psi_k'\left(A+\Psi_k\right)^{-1}\Psi_k\right)\mu_k - 2\mu_k'\Psi_k'\left(A+\Psi_k\right)^{-1}B - B'\left(A+\Psi_k\right)^{-1}B\right]\right\}, \\ &= \exp\left\{-\frac{1}{2}\left[\mu_k'\left(\Psi_k\left(A+\Psi_k\right)^{-1}A\right)\mu_k - 2\mu_k'\Psi_k'\left(A+\Psi_k\right)^{-1}B - B'\left(A+\Psi_k\right)^{-1}B\right]\right\}, \\ &= \exp\left\{-\frac{1}{2}\left(\left[\mu_k - \left(\Psi_k\left(A+\Psi_k\right)^{-1}A\right)^{-1}\Psi_k'\left(A+\Psi_k\right)^{-1}B\right]'\left[\Psi_k\left(A+\Psi_k\right)^{-1}A\right][\bullet]\right.\right. \\ &\qquad \left.\left. -B'\left[\left(A+\Psi_k\right)^{-1}\Psi_k\left(\Psi_k\left(A+\Psi_k\right)^{-1}A\right)^{-1}\Psi_k'\left(A+\Psi_k\right)^{-1} + \left(A+\Psi_k\right)^{-1}\right]B\right)\right\}, \\ &= \exp\left\{-\frac{1}{2}\left[\mu_k - A^{-1}B\right]'\left[\Psi_k\left(A+\Psi_k\right)^{-1}A\right][\bullet]\right\}\exp\left\{\frac{1}{2}B'A^{-1}B\right\}, \end{aligned}$$

where $A \equiv \widetilde{\alpha}'\Sigma\widetilde{\alpha}$ and $B \equiv \widetilde{\alpha}'\Sigma\widetilde{Y}_i^*$. Since the second exponential does not depend on Θ_i, nor on k, it can be absorbed into the factor of proportionality. This is why it vanishes in Equation (13).

CMS Discussion Paper 2010-03

For a complete list of the Discussion Papers published by the CMS please visit http://www.uni-konstanz.de/cms.

01 “The Effect of Grade Retention on School Outcomes: An Application of Doubly Robust Estimation Method” by Selver Derya Uysal, January 2010

02 “Can we help finance professionals to predict the Euribor rate?” by Fabian Krüger, Frieder Mokinski and Winfried Pohlmeier, February 2010

03 “Bayesian Inference for Factor Structure Models via Gibbs Sampling” by Rémi Piatek, February 2010