I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 11042

A Multinomial Probit Model with Latent Factors: Identification and Interpretation without a Measurement System

Rémi Piatek
Miriam Gensowski

SEPTEMBER 2017

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 11042

A Multinomial Probit Model with Latent Factors: Identification and Interpretation without a Measurement System

Rémi Piatek
University of Copenhagen

Miriam Gensowski
University of Copenhagen and IZA

SEPTEMBER 2017

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

A Multinomial Probit Model with Latent Factors: Identification and Interpretation without a Measurement System*

We develop a parametrization of the multinomial probit model that yields greater insight into the underlying decision-making process, by decomposing the error terms of the utilities into latent factors and noise. The latent factors are identified without a measurement system, and they can be meaningfully linked to an economic model. We provide sufficient conditions that make this structure identified and interpretable. For inference, we design a Markov chain Monte Carlo sampler based on marginal data augmentation. A simulation exercise shows the good numerical performance of our sampler and reveals the practical importance of alternative identification restrictions. Our approach can generally be applied to any setting where researchers can specify an *a priori* structure on a few drivers of unobserved heterogeneity. One such example is the choice of combinations of two options, which we explore with real data on education and occupation pairs.

JEL Classification: C11, C25, C35

Keywords: multinomial probit, latent factors, Bayesian analysis, marginal data augmentation, educational choice, occupational choice

Corresponding author:
Rémi Piatek
Department of Economics
University of Copenhagen
Øster Farimagsgade 5
1353 Copenhagen K
Denmark
E-mail: remi.piatek@econ.ku.dk

* This project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no 600207.

1 Introduction

The multinomial probit (MNP) model is a useful tool to estimate decision-making processes, especially when alternatives have correlated error terms. With an increasing number of alternatives, however, it becomes prohibitively difficult to estimate. The proliferation of parameters in the covariance matrix implies that it is not only computationally challenging, but also difficult to interpret, because a large, unstructured covariance matrix is utterly non-informative about the unobserved heterogeneity that drives choices. For example, taste shocks that apply to specific alternatives in a structured way are hidden in the error terms' noise.

To address the computational challenge, progress has been made by imposing some structure on the covariance matrix—for example by specifying a single latent factor (Geweke et al., 1994), a structured covariance matrix (Yai et al., 1997), zero restrictions determined stochastically in a data-driven way (Cripps et al., 2009), or autoregressive error terms (Bolduc, 1992). For panel data, researchers have specified multiple factors (Elrod and Keane, 1995), random effects (Nobile et al., 1997), and autoregressive error terms (Börsch-Supan et al., 1992). These approaches represent steps toward improved computation and interpretation, but many are tailored to panel data, and unfortunately the cross-sectional approaches do not make enough progress to interpret the covariance matrix. In some instances, identification questions remain, which would limit the inference that can be made about the underlying economic model.

To allow an interpretable latent structure, the MNP model can be combined with a structural equation model. In so-called ICLV models (integrated choice and latent variables), the underlying utilities are determined by latent factors that separate the unobserved heterogeneity into its different sources (Ben-Akiva et al., 2002; Daziano and Bolduc, 2013; Bhat and Dubey, 2014). While this approach is flexible (even if challenging to estimate, see Fu and Juan, 2017), it requires extra data to measure the latent factors. Yet, researchers often do not have such data available to extract the factors. At the same time, they may have a good idea *a priori* about how latent tastes or traits could map into the available choices. For example, in the study of simultaneous choices (Train, 2009), there is typically a clear structure of how choice-specific tastes relate to the different alternatives.

In the present paper, we develop a powerful parametrization of the MNP that decomposes the unobserved heterogeneity into latent factors, without requiring extra data to measure them. The fully identified and implied parsimonious latent structure, which links directly to an economic model, yields practical interpretability. The resulting implementation is computationally efficient compared to state-of-the-art methods.

Our methodological contribution is to develop a very general approach for cross-sectional data that relates a number of unobserved factors to the utilities of the alternatives. These (possibly correlated) factors can directly reflect an economic decision model and are not extracted from a measurement system, only assigned to the utilities through an allocation matrix. This structure reparametrizes the MNP to separate noise (idiosyncratic

error terms) from economic decision content (latent factors), and makes the covariance matrix manageable both for estimation and interpretation.

To ensure full economic interpretability, we place identification at the core of our analysis. New identification challenges arise because of the lack of a measurement system, as the factors are only allocated and not extracted. We provide several theoretical identification criteria, and also show how empirical identification is achieved in practice.

Our second contribution is computational. We develop an efficient approach relying on Bayesian methods for the inference of this factor structure model—its implementation in an `R` package is available from the authors upon request. Markov chain Monte Carlo (MCMC) methods have been successfully applied to the MNP model (McCulloch and Rossi, 1994, 2000; Nobile, 1998, 2000; McCulloch et al., 2000), with recent advances relying on marginal data augmentation that are even more efficient (Imai and van Dyk, 2005a, 2005b; Jiao and van Dyk, 2015). These latter methods introduce extra "working parameters" that cannot be identified from the data but serve to improve the sampler's convergence and mixing (van Dyk et al., 2001). We apply these techniques and construct suitable working parameters that build on our identification restrictions. Using synthetic and real data, we provide evidence that our sampler succeeds in decomposing the covariance structure between latent factors and noise. We show that our approach performs at least as well as the most recent developments for the standard MNP model (Imai and van Dyk, 2005a; Jiao and van Dyk, 2015), and that it outperforms them with respect to interpretability and economic content.

Our MNP with latent factors can be applied to any setting where unobserved tastes, effects, or features are present for some alternatives but not others, and where an allocation of these factors to the latent utilities can be specified beforehand. Throughout the paper, we use as a motivating example a model where agents choose simultaneously their occupation and level of education. In this joint decision process, the different alternatives reflect pairs of decision types, and are naturally correlated. The latent factors are specified at the education and occupation levels, so as to capture taste shocks associated with the options at each of these two decision types. This setup allows us to disentangle the channels of unobserved preferences, where the standard MNP model would remain silent, as the underlying factors would be hidden in the overall covariance matrix.

The paper is organized as follows. Section 2 lays out the specification and identification of our MNP model with latent factors. The identification problems at stake are thoroughly discussed, and formal proofs of identification are given. Section 3 introduces the Bayesian inferential procedure. We present marginal data augmentation methods and explain how we use them to construct an MCMC sampler that safeguards the identification of the model and that is efficient at the same time. Section 4 investigates the performance of the proposed sampler in a Monte Carlo experiment, and Section 5 presents the empirical application to the joint choice of occupation and education. Section 6 concludes.

2 Specification and identification of the multinomial probit model with latent factors

2.1 Model specification

Each agent $i = 1, \ldots, N$ takes a decision between $K+1$ alternative choices $D_i \in \{0, 1, \ldots, K\}$ by solving the following utility maximization problem:

$$D_i = \underset{k}{\operatorname{argmax}}\, U_{ik}, \tag{1}$$

$$U_{ik} = W_{ik}'\beta + \varepsilon_{ik}^{\star}, \tag{2}$$

for $k = 0, 1, \ldots, K$, where each utility U_{ik} is assumed to depend linearly on observed alternative-specific[1] characteristics W_{ik} through a vector of regression coefficients β.

To operationalize Eq. (1), distributional assumptions and identification restrictions are required on the error terms $\varepsilon_{ik}^{\star}$. Depending on these assumptions (e.g., logistic or normal distribution), this model could give rise to the well-known conditional logit, to the nested logit or to the multinomial probit model (see Train, 2009, for a review). In this paper, we introduce an alternative approach that assumes an *underlying latent structure* of the error terms. More specifically, each $\varepsilon_{ik}^{\star}$ is decomposed into J latent factors $\eta_{ij}^{\star}$ that are allocated to the latent utilities through an allocation matrix $\Gamma^{\star}$.

$$\varepsilon_{ik}^{\star} = \Gamma^{\star}\, \eta_i^{\star} + u_{ik}, \tag{3}$$

where $\Gamma^{\star}$ is a user-specified matrix of dimension $(K+1) \times J$, $\eta_i^{\star} = \left(\eta_{i1}^{\star}, \ldots, \eta_{iJ}^{\star}\right)'$ is the J-vector containing the latent factors, and u_{ik} is the $(K+1)$-vector of idiosyncratic errors.

For the sake of simplicity, all unobservables are assumed to be normally distributed:

$$\begin{pmatrix} \eta_i^{\star} \\ u_i \end{pmatrix} \mid W_i \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix} ; \begin{pmatrix} \Phi^{\star} & 0 \\ 0 & \Sigma^{\star} \end{pmatrix} \right), \tag{4}$$

with $\Sigma^{\star} = \operatorname{diag}\left(\sigma_0^2, \ldots, \sigma_K^2\right)$ and $W_i = (W_{i0}, W_{i1}, \ldots, W_{iK})'$. The matrix $\Phi^{\star}$ can either be full if the latent factors are assumed to be correlated, or a diagonal or block-diagonal matrix (i.e., with zero constraints) if they are uncorrelated, depending on the requirement of the underlying economic theory. The latent factors $\eta_i^{\star}$, as well as the error terms u_i, are assumed to be independent of the covariates, and the factors are assumed to be independent of the error terms. This specification results in the following covariance matrix for the unobserved part of the model:

$$\Omega^{\star} \equiv \operatorname{Var}(\varepsilon_i^{\star}) = \Gamma^{\star}\, \Phi^{\star}\, \Gamma^{\star\prime} + \Sigma^{\star}. \tag{5}$$

1. The covariates could also be individual-specific (i.e., fixed across alternatives), in which case the regression coefficients would vary across alternatives. We stick to the alternative-specific covariates in this notation, for the sake of simplicity, and because they help for identification in practice (Keane, 1992).

We use the notational convention that starred parameters refer to the unidentified version of the model, while later in the paper unstarred versions will denote the counterpart of these parameters in the identified version of the model. Identification issues and the restrictions they require will be discussed in the following sections.

We call the matrix $\Gamma^\star$ 'allocation matrix,' instead of 'factor loading matrix' in the terminology of the factor analysis literature, to prevent possible confusion. In our setup, this matrix is not estimated but fixed by the analyst. Usually, $\Gamma^\star$ will take the form of a binary matrix that determines the mapping of the $\eta^\star_{ij}$ into the latent utilities. We impose this fixed structure for two reasons. The first is an economic argument: In the applications we have in mind, the latent factors capture taste shocks that are associated with different features of the alternatives, and the interest lies in learning how these shocks are related. Specifying the mapping between latent utilities and factors usually comes naturally. Decomposing the covariance matrix $\Omega^\star$ is helpful for the interpretation of the decision process, as $\Omega^\star$ by itself is rather non-informative. The structure given by the allocation matrix induces parsimony, in the sense that fewer parameters describe the structure. Furthermore, the decomposition of $\Omega^\star$ separates the economic content in $\Phi^\star$ from noise in $\Sigma^\star$. The researcher can learn about the relative importance of these sources of unobserved heterogeneity, and even learn how the different factors in $\eta^\star_i$ are related to each other. For more intuition, we refer to the discussion in our example in Section 2.2.

The second reason for the fixed structure in $\Gamma^\star$ is a statistical one: Contrary to traditional latent factor models where the factors are extracted from multiple manifest variables, in our framework only the choice is observed—a single categorical variable—and the latent structure is obtained as the decomposition of the overall covariance matrix of the model. No extra information is available to measure the latent factors, contrary to integrated choice and latent variable (ICLV) models (Ben-Akiva et al., 2002; Bhat and Dubey, 2014). This complicates both the identification and the inference of the model. The fixed factor loadings structure we impose helps in this respect, as we will explain in the following sections.[2]

The informed reader may argue at this point that the standard MNP can easily accommodate any covariance matrix, including one generated by an underlying structure such as in Eq. (5). While this is true, we argue that our structural approach, with the decomposition in Eq. (3), provides a much clearer interpretation. The standard MNP would only be able to recover $\Omega^\star$ and ignore the underlying structure of this matrix on the right-hand side of the equation. Our approach not only addresses this potential structure, but uses it: its parsimonious structure facilitates the inference of larger models compared to the reduced-form version offered by the standard MNP model. As usual, this gain comes at a price: any misspecification in the structure of the latent part of the model may result in a deterioration of the results. How large the risk of misspecification is will depend on the application at hand. In the example we introduce below, we argue this risk to be rather

2. Estimating the factor loading matrix rather than fixing it would be possible in theory, but given the complexity of the task, we reserve this extension, and the investigation of its feasibility, for future research.

small. A comparison of the parameters of the overall covariance matrix $\Omega^\star$ with a regular MNP is always possible and advisable.

Finally, note that despite the fixed factor loading matrix of our model, our approach can easily be re-expressed as a standard factor model. Using the Cholesky decomposition of the covariance matrix of the factors, it is possible to define an alternative model with factor loading matrix $\widetilde{\Gamma}^\star \equiv \Gamma^\star(\Phi^\star)^{\frac{1}{2}}$ and covariance matrix the identity matrix, $\widetilde{\Phi}^\star = I_J$. This reparametrization would be observationally equivalent to Eq. (3) with the covariance structure of Eq. (5), and would correspond to a standard factor model. The factor loadings in $\widetilde{\Gamma}^\star$ would be estimated, with some restrictions implied on these parameters. In our framework, however, we prefer to fix the factor loading matrix and estimate the covariance matrix. The factor loading matrix has no particular meaning in the context of our choice model, while the covariance matrix is interesting to interpret, as it directly gives an idea of the importance and relatedness of the different taste shocks.

2.2 Example: A model of joint decisions for the study of education and occupation choices

Consider that individuals make their decisions about education and occupation simultaneously. Each available alternative is a combination of two decision types: one schooling level among N_S schooling alternatives, and one occupation among N_O occupations.[3] There is a total number of $N_S N_O$ joint alternatives. Two taste shocks are assumed to influence each level of decision: schooling-related shocks in $\eta_i^{\star S}$ for each available schooling level, and occupation-related shocks $\eta_i^{\star O}$ for the occupations.

To provide more intuition, we develop here the case with two schooling levels, $N_S = 2$, and three occupations, $N_J = 3$, corresponding to a total of 6 alternatives. The error term of the overall model in Eq. (3) can be decomposed as:

$$\varepsilon_i^\star = \Gamma^\star \eta_i^\star + u_i, \qquad \eta_i^\star = \begin{pmatrix} \eta_i^{\star O_1} \\ \eta_i^{\star O_2} \\ \eta_i^{\star O_3} \\ \eta_i^{\star S_1} \\ \eta_i^{\star S_2} \end{pmatrix}, \qquad \Gamma^\star = \begin{pmatrix} 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 \end{pmatrix}, \tag{6}$$

with $u_i = (u_{i1}, \ldots, \ldots, u_{i6})'$.[4]

The allocation matrix $\Gamma^\star$ determines which factors influence the different occupations and education levels. The elements of $\eta_i^\star$ can be interpreted as occupation/education-specific taste shocks, and $\Gamma^\star$ forms all of their possible combinations. The taste shocks

3. By "occupation," we do not mean a precise type of job, but rather a bundle of characteristics that make it possible to create broad groups of job types.

4. The general formulation of $\Gamma^\star$ would be $\begin{pmatrix} \iota_{N_S} \otimes I_{N_O} & I_{N_S} \otimes \iota_{N_O} \end{pmatrix}$, where the vector $\iota_N = (1, \ldots, 1)'$ contains N ones, while I_N is the $(N \times N)$-identity matrix, and $\otimes$ denotes the Kronecker product.

(factors) are allowed to correlate, with the following partitioned covariance matrix:

$$\Phi^\star = \begin{pmatrix} \Phi^\star_{OO} & \Phi^\star_{OS} \\ \Phi^\star_{SO} & \Phi^\star_{SS} \end{pmatrix},$$

where $\Phi^\star_{SO} = \Phi^{\star\prime}_{OS}$.

The need for different latent factors $\eta_i^{\star O}$ and $\eta_i^{\star S}$, as well as their possible non-zero correlation, arises from the economic content of this model. Typically, in a decision-making framework, agents are allowed to have unobserved taste shocks that make them more or less inclined towards each alternative. In the classic random-utility setup, these random components are usually added to the "explained" preference from observed individual characteristics or choice-specific attributes. These random components, however, always correspond to a specific *type* of choice. Since two *types* of choices are combined in our setup, we need to account for taste-shocks for each type of decision. For example, if a worker derives extra utility from working in a service sector job, we want to allow for this taste to apply whenever she considers a "service" occupation, regardless of whether she considers this occupation paired with a low education level or a high education level.

By modelling the choice as being from among *combinations* of an occupation and an education level, in a simultaneous choice setting, we merely assume that the expected utility of a given education level is influenced by the expected occupation, and vice versa. This resembles the anticipation of future utility from education in different occupations that is inherent in dynamic structural models such as Keane and Wolpin (1997) or Lee (2005). Agents do not ignore their expected occupation when deciding on education. For instance, someone interested in manual work will expect a low utility gain from engaging in a PhD program in astrophysics. Assuming that individuals already have a broad idea of the type of occupation they would like to have later when they choose their education means that econometrically, this choice can be treated as simultaneous.

2.3 Identification

The MNP model is notorious for being non-identified if no restrictions are imposed on its structure. Our version of this model with latent factors introduces additional challenges that need to be tackled appropriately. This section discusses the different identification problems at stake, and how we address them.

2.3.1 Location problem

The lack of natural location of the latent utilities—they can be shifted simultaneously without affecting their ordering, i.e., without changing the likelihood—creates a well-known identifiability problem (see Dansie, 1985; Bunch, 1991). The traditional solution is to define a base category, e.g., $k = 0$, and to work with the differences in utilities with respect to this baseline. Defining $Y_{ik} \equiv U_{ik} - U_{i0}$, the decision problem stated in Eq. (1)

can thus be re-expressed as:

$$D_i = \begin{cases} 0 & \text{if } \max(Y_i) < 0, \\ k & \text{if } \max(Y_i) = Y_{ik} \geq 0, \end{cases} \tag{7}$$

for $i = 1, \ldots, N$, where $\max(Y_i)$ is the maximal element of $Y_i = (Y_{i1}, \ldots, Y_{iK})'$. This representation is observationally equivalent to the original decision problem in Eq. (1) and solves the current identification issue.

The overall model in differenced form is obtained by pre-multiplying Eqs. (2) and (3) by the matrix Δ_K:

$$Y_i \equiv \Delta_K U_i = X_i \beta + \Gamma \eta_i + \omega_i, \tag{8}$$

where:

$$\Delta_K = \begin{pmatrix} -\iota_K & I_K \end{pmatrix}, \qquad X_i = \Delta_K W_i, \qquad \omega_i = \Delta_K u_i, \tag{9}$$

while for the latent factors, this first differentiation generally results in a reduction of their number. This reduction can be operationalized through two transformation matrices H and G:

$$\Gamma = \Delta_K \Gamma^\star H, \qquad \eta_i = G\, \eta_i^\star, \tag{10}$$

where η_i is a vector containing the P latent factors in their differenced form with respect to the factors appearing in the baseline utility U_{i0}. G is a matrix of dimension $(P \times J)$ that yet needs to be specified, and H is the corresponding $(J \times P)$-matrix, such that

$$\Gamma \eta_i = \Delta_K \Gamma^\star H G\, \eta_i^\star = \Delta_K \Gamma^\star \eta_i^\star. \tag{11}$$

Some conditions on both G and H, which will be discussed below, are required to make this transformation feasible.[5] The transformation in Eq. (10) looks more complicated than those of the covariates and the error terms in Eq. (9), because the latent factors driving the baseline utility U_{i0} may also influence other utilities, and cancel out selectively when the whole system is differentiated.

In most applications, the specification of G comes naturally, as it depends on how the latent factors cancel out in the first differentiation of the system. Our education-occupation joint decision example, continued in the following subsection, gives an illustration. In more sophisticated models, G and H might be less straightforward to specify. Proposition 2.1 provides a sufficient condition for G and H to be valid matrices for the required transformation, which can help specify them.

5. If G had a left generalized inverse H such that HG would be equal to the identity matrix, the problem would be trivial. This is not the case, unfortunately, because G is of dimension $(P \times J)$ and therefore only has at most $P < J$ linearly independent columns.

Proposition 2.1. *A sufficient condition for G and H to allow the transformation in Eq. (10), i.e., to fulfill the condition in Eq. (11), is that*

1. *G is made of P linearly independent rows of $\Delta_K \Gamma^\star$, where $P = \text{rank}\,(\Delta_K \Gamma^\star)$, or of any linear combination of the rows of $\Delta_K \Gamma^\star$ that provides P linearly independent rows,*

2. *H is the Moore-Penrose pseudoinverse of G.*

Proof. See Appendix A1. □

Because of the normality assumption made in Eq. (4) on the unobservables in the original model, the latent factors and error terms are also normally distributed in the differenced system, with following covariance matrices:

$$\Phi \equiv \text{Var}(\eta_i) = G\,\Phi^\star G', \tag{12}$$

$$\Sigma \equiv \text{Var}(\omega_i) = \Delta_K \Sigma^\star \Delta_K' = \sigma_0^2 \iota_K \iota_K' + \text{diag}(\sigma_1^2, \ldots, \sigma_K^2), \tag{13}$$

and the overall covariance matrix of the latent part of the differenced system is equal to:

$$\Omega = \Gamma\,\Phi\,\Gamma' + \Sigma. \tag{14}$$

Example: Differenced system in the education-occupation joint decision model. Since each latent utility is influenced by one schooling-specific effect and one occupation-specific effect, a natural transformation of the system in Eq. (6) is to subtract the two factors of the baseline utility (first occupation and first education factors) from the other factors, within each decision type. This can be obtained from the transformation in Eq. (10) by defining the matrix G as:

$$G = \begin{pmatrix} -1 & 1 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & -1 & 1 \end{pmatrix}, \qquad \underset{(P\times 1)}{\eta_i} = G\,\eta_i^\star = \begin{pmatrix} \eta_i^{\star O_2} - \eta_i^{\star O_1} \\ \eta_i^{\star O_3} - \eta_i^{\star O_1} \\ \eta_i^{\star S_2} - \eta_i^{\star S_1} \end{pmatrix}.$$

The general formulation for G that can be applied to any choice setting where two types of choices are combined is

$$G = \begin{pmatrix} \Delta_{(N_O-1)} & 0 \\ 0 & \Delta_{(N_S-1)} \end{pmatrix},$$

with Moore-Penrose pseudoinverse equal to:

$$H = \begin{pmatrix} \Delta^+_{(N_O-1)} & 0 \\ 0 & \Delta^+_{(N_S-1)} \end{pmatrix}, \qquad \Delta_N^+ = \begin{pmatrix} 0_{(1\times N)} \\ I_N \end{pmatrix} - \frac{\iota_{(N+1)} \iota_N'}{N+1}.$$

It can be verified that this choice of G and H fulfills the condition in Proposition 2.1. The allocation matrix corresponding to G and H in this general setting, where two choices are

made jointly, becomes

$$\underset{(K\times P)}{\Gamma} = \Delta_K \Gamma^\star H = \begin{pmatrix} I_{(N_O-1)} & 0_{[(N_O-1)\times(N_S-1)]} \\ \iota_{(N_S-1)} \otimes \begin{pmatrix} 0_{[1\times(N_O-1)]} \\ I_{(N_O-1)} \end{pmatrix} & I_{(N_S-1)} \otimes \iota_{N_O} \end{pmatrix},$$

where $P = (N_O - 1) + (N_S - 1)$. More specifically, in our example with $P = 3$ factors, this matrix is

$$\Gamma = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix}.$$

Finally, the covariance matrix of the latent factors is the following partitioned matrix:

$$\text{Var}(\eta_i) = \begin{pmatrix} \Delta_{(N_O-1)} \Phi^\star_{OO} \Delta'_{(N_O-1)} & \Delta_{(N_O-1)} \Phi^\star_{OS} \Delta'_{(N_O-1)} \\ \Delta_{(N_S-1)} \Phi^\star_{SO} \Delta'_{(N_S-1)} & \Delta_{(N_S-1)} \Phi^\star_{SS} \Delta'_{(N_S-1)} \end{pmatrix}.$$

Necessary and sufficient conditions for identification. The identification of Φ and Σ from Ω, using the mapping defined in Eq. (14), hinges on two conditions. First, the number of parameters in the structural system with Φ and Σ should not exceed the number of available linear equations, i.e., the number of free parameters in Ω in the reduced-form model. Hence the following condition:

$$\frac{P(P+1)}{2} + K + 1 \quad \leq \quad \frac{K(K+1)}{2},$$

which is fulfilled for all $K > P + 1$.[6] This condition on the number of free parameters relative to the number of available equations is, however, only necessary and not sufficient.

An additional condition is required to achieve identification. For any pairs of matrices $\{\Phi, \Sigma\}$ and $\{\check{\Phi}, \check{\Sigma}\}$, such that there exists covariance matrices Ω and $\check{\Omega}$ fulfilling Eq. (14) for both pairs, i.e., $\Omega = \Gamma \Phi \Gamma' + \Sigma$ and $\check{\Omega} = \Gamma \check{\Phi} \Gamma' + \check{\Sigma}$, the overall covariance matrices can only be equal, $\Omega = \check{\Omega}$, if and only if $\Phi = \check{\Phi}$ and $\Sigma = \check{\Sigma}$. We obtain this result by making two assumptions on the structure of the allocation matrix Γ in the differenced system, which are stated in Assumption 2.1.

Assumption 2.1. *The allocation matrix Γ in the differenced system is such that:*

1) It is full rank, i.e., $\text{rank}\,(\Gamma) = P$.

6. In our example, this condition is fulfilled for any $N_S \geq 2$ and $N_O \geq 2$, where at least one of the two inequalities is strict. A model with $N_O = N_S = 2$ would therefore not be identified without any additional restrictions—at least one restriction on Φ or Σ would be required.

2) Every row of Γ is a linear combination of some (or all) of its other rows.

The second condition corresponds to a row-deletion property: Any row of Γ can be deleted without reducing the full rank of the matrix. This property is similar to identification requirements relying on rank conditions that are traditionally used in standard factor analysis, such as in Anderson and Rubin (1956, see, e.g., Theorem 5.1). They are slightly different in our framework, because we do not deal with the identification of the factor loading matrix—which is fixed in our model—but with the identification of the covariance matrix of the factors—which is fixed to the identity matrix in standard factor analysis. Therefore, we state and prove the full conditions for the identification of our model, for the sake of completeness.

Assumption 2.1, combined with the necessary condition on the number of latent factors relative to the number of alternatives ($K > P+1$), is sufficient for identification, as stated in the following proposition:

Proposition 2.2. *If the allocation matrix Γ defined in Eq. (10) satisfies Assumption 2.1, in a model where $K > P + 1$, then the covariance matrix of the latent factors Φ and the idiosyncratic variances $\sigma_0^2, \sigma_1^2, \ldots, \sigma_K^2$ are identified from the overall covariance matrix Ω.*

Proof. The proof relies on a rank condition, see details in Appendix A2. □

Although the linear function corresponding to Eq. (14) is bijective, the proof of Proposition 2.2 only uses its injectivity to show that Φ and $\sigma_0^2, \sigma_1^2, \ldots, \sigma_K^2$ are identified from Ω. The surjectivity of this function cannot be exploited: If all covariance matrices Φ and all variances $\sigma_0^2, \sigma_1^2, \ldots, \sigma_K^2$ yield a matrix Ω that is positive semi-definite, the reverse is not true and not all covariance matrices Ω correspond to a positive semi-definite matrix Φ and strictly positive variances $\sigma_0^2, \sigma_1^2, \ldots, \sigma_K^2$. This remark has important implications for the inference of the model. Although in theory it would be possible to work with the reduced-form model that only involves Ω, and to retrieve the corresponding parameters Φ and $\sigma_0^2, \sigma_1^2, \ldots, \sigma_K^2$, this approach would require an estimation under constraint to guarantee the positive semi-definiteness and positiveness of the corresponding parameters. In the Bayesian framework used in this paper, this would imply sampling from non-standard truncated distributions. Additionally, this would introduce complications in the interpretation of the prior distribution—the prior would be specified on Ω, and would induce priors on Φ and Σ that would not be straightforward to derive and might have odd shapes. For these reasons, it turns out to be easier to work with the structural form model using the right-hand side of Eq. (14), for the sake of inference and interpretation of the model.

In the remainder of the paper, we work with the differenced system in Eq. (8). To simplify text and notation, we refer to the *differenced latent factors* simply as the *latent factors*, and similarly for the utilities and the error terms.

Coming back to our example, it can easily be shown that in this case Assumption 2.1 holds by reordering the rows of Γ such that each row containing a single 1 appears on top

to form the identity matrix:

$$\underset{(K\times P)}{\overline{\Gamma}} = \begin{pmatrix} I_P \\ \iota_{(N_S-1)} \otimes I_{(N_O-1)} \quad I_{(N_S-1)} \otimes \iota_{(N_O-1)} \end{pmatrix} \equiv \begin{pmatrix} I_P \\ \overline{\Gamma}_2. \end{pmatrix} \tag{15}$$

Since the top block is the identity matrix, $\overline{\Gamma}$ can be put in reduced row echelon form, which implies it is full rank. Given that each row and each column of the lower block $\overline{\Gamma}_2$ contains two nonzero elements, respectively, and the upper block is the identity matrix, any row of $\overline{\Gamma}$ can be obtained from elementary operations on two of its other rows. This fulfills the second point of Assumption 2.1.

2.3.2 Scaling problem

The second well-known identification problem arises because the latent utilities U_{ik} have no natural scale—they can all be multiplied by a positive constant without changing their ordering. Therefore, one restriction that sets the scale of the utilities is required to achieve identification, even in the differenced system in Eq. (8).

In the MNP framework, this identification problem is commonly solved by fixing one of the diagonal elements of the covariance matrix of the error terms to a constant—usually, the first variance is set to one (see McCulloch and Rossi, 1994; McCulloch et al., 2000). Alternatively, Burgette and Nordheim (2012) impose a restriction on the trace of the covariance matrix. This restriction makes the model symmetric with respect to the choice of the baseline level, and exhibits computational advantages over the standard restriction of the MNP model.

Given the latent structure of our model, many different strategies can be implemented to address this identification problem. We consider the following restrictions, for a user-defined constant $c \in \mathbb{R}^+$:

Table 1: Identification restrictions, four strategies

	Idiosyncratic variances	Overall covariance matrix
Element restriction	**R1a** $\sigma_1^2 = c$	**R2a** $\Omega_{[1,1]} \equiv \Phi_{[1,1]} + \sigma_0^2 + \sigma_1^2 = c$
Trace restriction	**R1b** $\sum_{k=0}^{K} \sigma_k^2 = c$	**R2b** $\text{tr}(\Phi) + \sum_{k=0}^{K} \sigma_k^2 = c$

All these restrictions prevent the utilities from being rescaled, but they operate through different channels. First, the two R1* restrictions only fix the variance(s) of the error term(s) in Σ, while the R2* conditions impose the restriction on the overall covariance matrix of the latent part of the model, Ω. The latter ones are therefore equivalent to the way identification is usually achieved in the MNP model. Second, the two R*a restrictions require the analyst to select one baseline utility to impose the restriction (in this case, the first one), whereas the two R*b restrictions impose the restriction on a combination

of all the parameters affected by the scaling problem. It resembles the trace restriction of Burgette and Nordheim (2012) and does not create any asymmetry between the utilities because of the choice of the restriction.

These four identification strategies are innocuous for the interpretation of the model—they only set the scale of the utilities in different ways—and they do not affect the general identification of the model stated in Proposition 2.2.[7] However, they have very different implications for the inference of the model. Only the first one, `R1a`, can be implemented in a simple way, as it only restricts a single parameter. All others put the restriction on a combination of several parameters, which makes the whole estimation problem far less trivial. Besides this practical issue, they may also imply different properties of the estimator. In the Bayesian setting adopted in this paper, this translates into different features of the sampler, as well as different prior distributions on the parameters of the model. It is crucial to grasp these subtleties to fully understand how to carry out inference and to interpret the results.

3 Marginal data augmentation methods for the inference of the identified model

We now present the algorithm used for inference that guarantees the identification conditions just laid out. This algorithm has been implemented as an `R` package, so that interested researchers can apply the proposed method conveniently.[8]

Without the identification restriction, it would be straightforward to make inference on our model. Standard data augmentation methods could be used to simulate the unobserved latent factors and latent utilities (Tanner and Wong, 1987), and no complications would arise in the implementation of a simple Gibbs sampler. The restriction required for identification, unfortunately, completely changes the setup, as no prior distribution that fulfills this identification condition can be found (except in the `R1a` case). Therefore, there is a dichotomy between two versions of the model: the identified version we are interested in but that is impractical to handle on the one hand, and the non-identified version in which a sampler can be derived but that is difficult to interpret because of the lack of identification on the other hand. In this section, we propose to exploit this dichotomy through the use of Marginal Data Augmentation methods. These enable us to work effortlessly with an unidentified model, then move back to an identified and therefore interpretable version of this model.

7. As a general result, applying linear restrictions on the parameters of an identified model does not affect its identification. In all four identification strategies, σ_1^2 can be expressed as a linear function of the constant c and of (some or all of) the remaining parameters $\text{tr}(\Phi), \sigma_0^2, \sigma_2^2, \ldots, \sigma_K^2$. Dropping this constrained parameter reduces the number of parameters by one, and at the same time the rank of the system of linear equations by one, but the overall system remains full rank.

8. Package available from the authors upon request.

3.1 Principle

Marginal Data Augmentation (MDA) methods have been introduced by Meng and van Dyk (1999) and van Dyk et al. (2001), and developed in parallel with Parameter-Expansion methods (Liu et al., 1998; Liu and Wu, 1999). These approaches started from the observation that one can dramatically boost the convergence and mixing of the MCMC sampler by relaxing model restrictions, notably through the introduction of extra parameters that cannot be identified from the data. This surprising property has been extensively and successfully applied to models where convergence and mixing can be problematic, such as in latent variable models, where the proliferation of latent variables usually makes parameter autocorrelations very high. A by-product of MDA methods is that they allow to deal with constrained models that would otherwise be very hard or even impossible to handle. They have been successfully implemented to the standard multinomial probit by Imai and van Dyk (2005a) and Jiao and van Dyk (2015).

To fix ideas and understand how MDA proceeds, let us express the likelihood function of the model, which depends on the model parameters β, Φ and Σ, in two different ways:

$$\begin{aligned}
\mathcal{L}(\beta, \Phi, \Sigma \mid D, X) &\propto p(D \mid X, \beta, \Phi, \Sigma), \\
&\propto \iint p(D, Y, \eta \mid X, \beta, \Phi, \Sigma) dY\, d\eta, \qquad (16) \\
&\propto \iint \left\{ \int p(D, Y, \eta \mid X, \beta, \Phi, \Sigma, \alpha^2) p(\alpha^2 \mid \Phi, \Sigma) d\alpha^2 \right\} dY\, d\eta. \qquad (17)
\end{aligned}$$

In Eq. (16), the likelihood is *augmented* with the latent utilities Y and latent factors η. This is the traditional Data Augmentation approach (Tanner and Wong, 1987), which explicitly incorporates the latent variables into the model to facilitate sampling. In Eq. (17), an extra parameter α^2 is introduced and averaged out of the likelihood over its conditional distribution $p(\alpha^2 \mid \Phi, \Sigma)$. This is the *Marginal* Data Augmentation approach. The extra parameter α^2, commonly called *working parameter*, cannot be identified by the data, i.e., its introduction into the model does not alter the likelihood, such that $\mathcal{L}(\beta, \Phi, \Sigma \mid D, X) = \mathcal{L}(\beta, \Phi, \Sigma, \alpha^2 \mid D, X)$.

In our model, an obvious candidate for α^2 is the scale of the latent utilities. As discussed previously in Section 2.3, the utilities have no natural scale. Therefore there exists an infinite number of parameters $\alpha^2 \in \mathbb{R}^+$ that can be used to multiply the utilities while leaving the likelihood of the model unchanged. We use this property for the implementation of MDA, similar to Imai and van Dyk (2005a) and Jiao and van Dyk (2015).

The working parameter and the resulting model expansion have to be chosen in such a way that there exists a one-to-one mapping between the parameters of the expanded model and those of the original model. This bijection ensures that it is possible to move between the two versions of the model in a unique way, and always allows to move back to the identified model.

Given the four identification strategies outlined in Section 2.3, four corresponding

working parameters can be considered. Each identification restriction implies different prior distributions on the idiosyncratic variances and the covariance matrix of the latent factors. Each calls for a different sampling scheme, and might perform differently in practice. We will discuss these differences and illustrate them through simulations.

3.2 Model expansion

The model in Eq. (8), assumed to be identified thanks to a restriction imposed on the covariance matrix of its latent part, can be expanded by rescaling all the utilities by an auxiliary parameter $\alpha \in \mathbb{R}^+$:

$$\alpha Y \equiv \widetilde{Y}_i = X_i\widetilde{\beta} + \Gamma\widetilde{\eta}_i + \widetilde{\omega}_i, \qquad \begin{aligned} \widetilde{\eta}_i &\sim \mathcal{N}\left(0; \widetilde{\Phi}\right), \\ \widetilde{\omega}_i &\sim \mathcal{N}\left(0; \widetilde{\Sigma}\right), \end{aligned} \tag{18}$$

with:

$$\widetilde{\beta} \equiv \alpha\beta, \qquad \widetilde{\eta}_i \equiv \alpha\eta_i, \qquad \widetilde{\omega}_i \equiv \alpha\omega_i, \qquad \widetilde{\Phi} \equiv \alpha^2\Phi, \qquad \widetilde{\Sigma} \equiv \alpha^2\Sigma, \tag{19}$$

where tildes are used on the parameters of the expanded model to distinguish them from their counterparts in the original (identified) model. This transformation does not affect the observational rule in Eq. (7), so expanded utilities $\widetilde{Y}_i$ can be used in place of Y_i.

The auxiliary parameter α^2 qualifies as a *working parameter* for MDA: it is not identifiable from the data in the unrestricted model, and it allows to define a one-to-one mapping between the parameters of the two versions of the model, as expressed in Eq. (19). The working parameter has to be expressed differently in the four identification strategies to ensure that the restrictions defined in Table 1 are fulfilled, see Table 2.

Table 2: Working parameters in the four identification strategies

R1a:	$\alpha^2 = \widetilde{\sigma}_1^2/c,$	**R2a:**	$\alpha^2 = \left(\widetilde{\Phi}_{[1,1]} + \widetilde{\sigma}_0^2 + \widetilde{\sigma}_1^2\right)/c,$
R1b:	$\alpha^2 = \left(\sum_{k=0}^K \widetilde{\sigma}_k^2\right)/c,$	**R2b:**	$\alpha^2 = \left(\mathrm{tr}\left(\widetilde{\Phi}\right) + \sum_{k=0}^K \widetilde{\sigma}_k^2\right)/c,$

where $c \in \mathbb{R}^+$ is the user-defined value of the restriction.

Now that we have defined the relationship between the parameters of the identified model and those of its expanded version, we can derive the prior distribution implied on the former ones when a proper prior is specified on the latter ones.

3.3 Working prior distribution

The implementation of MDA requires us to average the likelihood over a prior distribution for the working parameter $p(\alpha^2 \mid \Phi, \Sigma)$ in Eq. (17). For a given prior distribution on the parameters of the expanded model $\widetilde{\Phi}$ and $\widetilde{\sigma}_k^2$, each identification strategy implies a different

prior distribution on the working parameter α^2 and on the corresponding parameters Φ and σ_k^2 in the identified model. It is important to understand how the prior distribution of the unrestricted parameters is related to the prior of the restricted parameters and of the working parameters to implement MDA.

3.3.1 Prior for identification restrictions on idiosyncratic variances

In schemes R1a and R1b, the covariance matrix of the latent factors Φ is left unconstrained, and only the idiosyncratic variance(s) are restricted to set the scale of the latent utilities. Proposition 3.1 and Corollaries 3.1 and 3.2 provide the analytical results for the conditional prior of the working parameter α^2, as well as the prior distribution induced on $\sigma_0^2, \sigma_1^2, \ldots, \sigma_K^2$ in the identified model.

Proposition 3.1 (**Conditional prior distribution of α^2 in scheme R1a**). *Assuming inverse-Gamma prior distributions on the idiosyncratic variances in the expanded model,*

$$\widetilde{\sigma}_k^2 \sim \mathcal{G}^{-1}(a_0; b_0)\,, \qquad a_0, b_0 > 0, \tag{20}$$

for $k = 0, \ldots, K$, the variable transformation used in scheme R1a implies the following conditional prior distribution for the working parameter:

$$\alpha^2 \mid \sigma_0^2, \ldots, \sigma_K^2 \sim \mathcal{G}^{-1}\left(a_0(K+1); b_0 \sum_{k=0}^{K} \frac{1}{\sigma_k^2}\right). \tag{21}$$

Corollary 3.1 (**Marginal prior distribution of Σ in scheme R1a**). *Given the priors in Eqs. (20) and (21), the marginal prior distribution of the idiosyncratic variances in scheme R1a is proportional to:*

$$p(\sigma_0^2, \ldots, \sigma_K^2) \propto \left(\sum_{k=0}^{K} \frac{1}{\sigma_k^2}\right)^{-a_0(K+1)} \prod_{k=0}^{K} (\sigma_k^2)^{-a_0-1}\, \mathbb{1}\{\sigma_1^2 = c\}\,. \tag{22}$$

Proof. The transformation of random variables provides:

$$\begin{aligned} p(\alpha^2, \sigma_0^2, \ldots, \sigma_K^2) &= \mathcal{J}_{(\widetilde{\sigma}_0^2, \ldots, \widetilde{\sigma}_K^2) \to (\alpha^2, \sigma_0^2, \ldots, \sigma_K^2)}\, p(\widetilde{\sigma}_0^2, \ldots, \widetilde{\sigma}_K^2), \\ &\propto \left(\alpha^2\right)^K \prod_{k=0}^{K} (\widetilde{\sigma}_k^2)^{-a_0-1} \exp\left\{-\frac{b_0}{\widetilde{\sigma}_k^2}\right\} \mathbb{1}\{\sigma_1^2 = c\}\,, \\ &\propto \left(\alpha^2\right)^{-a_0(K+1)-1} \exp\left\{-\frac{b_0}{\alpha^2} \sum_{k=0}^{K} \frac{1}{\sigma_k^2}\right\} \prod_{k=0}^{K} (\sigma_k^2)^{-a_0-1}\, \mathbb{1}\{\sigma_1^2 = c\}\,, \end{aligned} \tag{23}$$

where the Jacobian of the transformation $\mathcal{J}_{(\cdot)\to(\cdot)}$ adds a factor of $\left(\alpha^2\right)^K$,[9] and $\mathbb{1}\{\sigma_1^2 = c\}$ is the indicator function that is equal to one if the corresponding condition is fulfilled, to zero otherwise. The kernel of an inverse-Gamma distribution can be extracted from the

9. See proof in Appendix B1.

last expression, proving Proposition 3.1. Corollary 3.1 is obtained by integrating out α^2 from Eq. (23), such that $p(\sigma_0^2, \ldots, \sigma_K^2) = \int p(\alpha^2, \sigma_0^2, \sigma_2^2, \ldots, \sigma_K^2)\, d\alpha^2$. □

Corollary 3.2 (Prior distributions of α^2 and Σ in scheme R1b). *In scheme R1b, assuming the same prior on $\widetilde{\sigma}_k^2$ as in Eq. (20), for $k = 0, \ldots, K$, implies that the conditional prior distribution of α^2 and the marginal prior distribution of $\sigma_0^2, \ldots, \sigma_K^2$ are the same as in Eq. (21) and Eq. (22), up to the indicator function that should be replaced by $\mathbb{1}\left\{\sum_{k=0}^K \sigma_k^2 = c\right\}$ in the latter equation.*

Proof. This result is a direct implication of the fact that the Jacobians of the two transformations R1a and R1b are identical, see Appendices B1 and B2. □

The kernel in Eq. (22) does not correspond to a known distribution and cannot be further factorized into a product of kernels for the different parameters σ_k^2. This result can be understood intuitively, as the idiosyncratic variances are all bound together *a priori* because of the identifying restriction—setting the scale of the first utility automatically sets the scales of the other utilities. Fortunately, it is straightforward to simulate this constrained prior distribution to get an idea of its shape. This can be done by sampling the parameters from the working prior in Eq. (20), and rescaling them appropriately to guarantee that the restriction is fulfilled.

The analogy between schemes R1a and R1b is comparable to what Burgette and Nordheim (2012) find for the MNP model, where the trace restriction they propose implies the same working prior distribution as for the original MNP model developed by Imai and van Dyk (2005a). In practice, this result is very convenient, as the same sampling scheme can be designed for both R1a and R1b. Only the step where the working parameter is retrieved from the sampled values of the idiosyncratic variances will differ.

3.3.2 Prior for identification restrictions on the overall covariance matrix Ω

The case of schemes R2a and R2b is more complicated, as the restriction now involves both the covariance matrix of the latent factors $\widetilde{\Phi}$ and the idiosyncratic variances $\widetilde{\sigma}_0^2, \ldots, \widetilde{\sigma}_K^2$. The prior distributions in the expanded model and the implied prior distributions in the identified model are summarized in Proposition 3.2 and Corollaries 3.3 and 3.4.

Proposition 3.2 (Conditional prior distribution of α^2 in scheme R2a). *Assuming that the parameters of the expanded model follow an inverse-Wishart distribution and inverse-Gamma distributions a priori, respectively,*

$$\widetilde{\Phi} \sim \mathcal{W}^{-1}(\nu_0;\, t_0 S_0)\,, \qquad \nu_0 \geq P, S_0 > 0 \text{ (pos. def.)} \tag{24}$$

$$\widetilde{\sigma}_k^2 \sim \mathcal{G}^{-1}(a_0;\, t_0 b_0)\,, \qquad a_0, b_0, t_0 > 0, \tag{25}$$

for $k = 0, \ldots, K$, the conditional prior distribution of the working parameter α^2 is the

following scaled inverse chi-squared distribution:

$$\alpha^2 \mid \Phi, \Sigma \sim t_0 \left(tr\big(S_0(\Phi)^{-1}\big) + 2b_0 \sum_{k=0}^{K} \frac{1}{\sigma_k^2} \right) / \chi^2_{(\nu_0 P + 2a_0(K+1))}. \tag{26}$$

Corollary 3.3 (**Marginal prior distribution of Φ and of Σ in scheme R2a**). *Given the priors in Eqs. (24) to (26), the joint prior distribution of Φ and Σ in the identified model is proportional to:*

$$\begin{aligned} p(\Phi, \Sigma) \propto & \left(tr\big(S_0(\Phi)^{-1}\big) + 2b_0 \sum_{k=0}^{K} \frac{1}{\sigma_k^2} \right)^{-(\nu_0 P + 2a_0(K+1))/2} \\ & \times |\Phi|^{-\frac{\nu_0+P+1}{2}} \prod_{k=0}^{K} \left(\sigma_k^2\right)^{-a_0-1} \mathbb{1}\left\{\Phi_{[1,1]} + \sigma_0^2 + \sigma_1^2 = c\right\}. \end{aligned} \tag{27}$$

Proof. From the variable transformation expressed in Eq. (19), the joint prior distribution of the corresponding parameters in the restricted model and of the working parameter is derived as follows:

$$\begin{aligned} p(\Phi, \Sigma, \alpha^2) &= \mathcal{J}_{(\widetilde{\Phi},\widetilde{\Sigma})\to(\Phi,\Sigma,\alpha^2)}\, p(\widetilde{\Phi}, \widetilde{\Sigma}), \\ &\propto \left(\alpha^2\right)^{\frac{P(P+1)}{2}+K} \left|\widetilde{\Phi}\right|^{-\frac{\nu_0+P+1}{2}} \exp\left\{-\frac{1}{2}\mathrm{tr}\left(t_0 S_0(\widetilde{\Phi})^{-1}\right)\right\} \\ &\quad \times \prod_{k=0}^{K} \left(\widetilde{\sigma}_k^2\right)^{-a_0-1} \exp\left\{-\frac{t_0 b_0}{\widetilde{\sigma}_k^2}\right\} \mathbb{1}\left\{\Phi_{[1,1]} + \sigma_0^2 + \sigma_1^2 = c\right\}, \\ &\propto \left(\alpha^2\right)^{-(\nu_0 P + 2a_0(K+1))/2-1} \exp\left\{-\frac{t_0}{2\alpha^2}\left[\mathrm{tr}\big(S_0(\Phi)^{-1}\big) + 2b_0 \sum_{k=0}^{K} \frac{1}{\sigma_k^2}\right]\right\} \\ &\quad \times |\Phi|^{-\frac{\nu_0+P+1}{2}} \prod_{k=0}^{K} \left(\sigma_k^2\right)^{-a_0-1} \mathbb{1}\left\{\Phi_{[1,1]} + \sigma_0^2 + \sigma_1^2 = c\right\}, \end{aligned} \tag{28}$$

where $\mathcal{J}_{(\widetilde{\Phi},\widetilde{\Sigma})\to(\Phi,\Sigma,\alpha^2)} \propto \left(\alpha^2\right)^{\frac{P(P+1)}{2}+K}$ is the Jacobian of the corresponding transformation (see proof in Appendix B3). The kernel of a scaled inverse chi-squared distribution can be extracted from Eq. (28), proving Proposition 3.2. Corollary 3.3 is obtained by integrating out α^2 from Eq. (28), such that $p(\Phi, \Sigma) = \int p(\Phi, \Sigma, \alpha^2) d\alpha^2$. □

Corollary 3.4 (**Prior distributions of α^2, Φ and Σ in scheme R2b**). *In R2b, assuming the same priors as in Eqs. (24) and (25) for $\widetilde{\Phi}$ and $\widetilde{\sigma}_k^2$, $k = 0, \ldots, K$, implies the same conditional prior distribution for the working parameter α^2 as in Eq. (26) and the same marginal prior distribution for the parameters of the identified model as in Eq. (27), up to the indicator function at the end of the latter that is equal to $\mathbb{1}\left\{tr(\Phi) + \sum_{k=0}^{K} \sigma_k^2 = c\right\}$.*

Proof. This result comes from the fact that the Jacobians of the two transformations R2a and R2b are identical, see Appendices B3 and B4. □

The scale matrix of the inverse-Wishart distribution and the scale parameter of the inverse-Gamma distributions depend on a common parameter t_0 in Eqs. (24) and (25). This parameter ensures that the two parts of the unobservables of the model are scaled similarly in the expanded model, and also allows to simplify calculations. Note that this parameter appears in the conditional distribution of the working parameter in Eq. (26), but not in the joint distribution of Φ and Σ in Eq. (27). Therefore, t_0 is a *tuning parameter* that controls to which degree the unobservables of the model are inflated in the expanded model, but does not affect the resulting prior distribution of the corresponding parameters in the restricted model.

As previously, the kernel in Eq. (27) cannot be further factorized into the product of two known kernels, because the identification restriction generates prior dependence between the covariance matrix Φ and the idiosyncratic variances σ_k^2. This prior can, however, also be simulated to get an idea of its shape.

3.3.3 Prior distribution of the remaining parameters

The regression parameters are affected by the rescaling (see Eq. (19)) but are not directly connected to the working parameter. As a consequence, it is possible to specify their prior distribution in the identified model, and to derive their implied prior in the expanded model conditional on the working parameter. Using a normal prior provides:

$$\beta \sim \mathcal{N}(0;\, B_0)\,, \qquad\qquad \widetilde{\beta} \mid \alpha^2 \sim \mathcal{N}\big(0;\, \alpha^2 B_0\big)\,, \tag{29}$$

In schemes `R1a` and `R1b` the covariance matrix of the latent factors, Φ, is also not directly connected to the working parameter, but still affected by the transformation. For these two schemes, we assume an inverse-Wishart distribution a priori, resulting in the same type of prior in the expanded model:

$$\Phi \sim \mathcal{W}^{-1}(\nu_0;\, S_0)\,, \qquad\qquad \widetilde{\Phi} \mid \alpha^2 \sim \mathcal{W}^{-1}\big(\nu_0;\, \alpha^2 S_0\big)\,. \tag{30}$$

3.4 Sampling scheme: MDA and partial collapsing

With the prior distribution of the working parameter and of the model parameters in hand, both in the expanded model and in the identified model, we can design a sampling scheme that implements the MDA approach. Our algorithm updates the parameters and the latent variables of the model iteratively according to the steps described below, where the working parameter α^2 is sampled alongside to allow the marginal data augmentation procedure to operate.

The sampler is presented in Algorithm 1. Each posterior distribution implicitly conditions on the observed data decision D and on covariates X, and the conditioning set always includes the most up-to-date values of the parameters and latent variables. Some steps contain intermediate values of some parameters that are immediately discarded—e.g., for the working parameters these intermediate steps are denoted $\alpha^{(a)}$, $\alpha^{(b)}$, and $\alpha^{(c)}$.

The corresponding conditional distributions are provided in Appendix C.

The covariance matrix Σ of the error terms in the differenced system has a particular structure[10] that makes it impossible to sample the variances σ_k^2 directly using the standard Gibbs sampler. Instead, we rely on data augmentation methods (Tanner and Wong, 1987) and simulate the error term $\widetilde{u}_0$ of the baseline utility in the expanded model. This simple one-factor error structure approach was proposed by Geweke et al. (1994, Section V) and is straightforward to implement.

This MCMC sampler has a number of interesting features that are worth pointing out. It is a non-standard MCMC sampler that combines (marginal) data augmentation techniques, a partial collapsing and a partial marginalization of the Gibbs sampler (van Dyk and Park, 2008, 2009), to generate a Markov chain with improved mixing properties and better convergence.

In step 1, the working parameter is sampled from its conditional prior distribution that depends on Φ and Σ in the identified model, as no other information is available at this stage to move to the expanded model. This is done using the results of Section 3.3.1 or Section 3.3.2, depending on the chosen identification strategy. Steps 1 and 2 are carried out conditional on the covariance matrices Φ and Σ, but not on the latent variables $\widetilde{\eta}$ and $\widetilde{u}_0$. Integrating out these latent variables in some steps of the sampler, while explicitly drawing them and conditioning on them in other steps (e.g., in steps 3 and 4), is allowed in the framework of a *partially collapsed Gibbs sampler* (van Dyk and Park, 2008, 2009). As emphasized by these authors, partially collapsing the Gibbs sampler must be done with care, as it may alter the stationary distribution of the Markov chain. Particularly, only intermediate quantities that are *not* conditioned upon in subsequent steps of the sampler can be marginalized and trimmed safely.[11] This is the case here: Since $\widetilde{\eta}$ and $\widetilde{u}_0$ do not appear in any conditioning set until they are sampled in step 3, they can be marginalized and trimmed in the first two steps of the sampler.

The implicit goal of step 2 is to sample β from $p(\beta \mid \widetilde{Y}, \Phi, \Sigma)$. This is done by sampling jointly the regression parameters and the working parameter from $p(\beta, \alpha^2 \mid \widetilde{Y}, \Phi, \Sigma)$, which is the same as sampling from $p(\widetilde{\beta}, \alpha^2 \mid \widetilde{Y}, \Phi, \Sigma)$ and transforming $\beta = \widetilde{\beta}/\alpha$. Importantly, β and α^2 need to be sampled simultaneously, so that the marginal data augmentation procedure does not distort the prior distribution of the regression parameters. In step 3, the working parameter is not sampled jointly with the latent variables $\widetilde{\eta}$ and $\widetilde{u}_0$, but rather conditioned upon. The fact that the working parameter is not sampled in each step of the sampler implies that we are dealing with a *partially marginalized Gibbs sampler* (see van Dyk, 2010, Section 3.2). This is a valid step in this framework, as $\widetilde{\eta}$ and $\widetilde{u}_0$ only need to be sampled in the expanded model to then allow the sampling of Φ and Σ in step 4. The values of these latent variables in the identified model are of no particular interest, and they are not used in the first two steps of the sampler. Hence, their trimming is possible

10. See Eq. (13).

11. I.e., they can be moved from the conditioning set to the set of parameters or latent variables being sampled, and then discarded from the sampling scheme for these steps.

Algorithm 1 MCMC Sampler

step 0: Set $t \leftarrow 0$. Initialize model with starting values $\beta^{(0)}$, $\Phi^{(0)}$, $\Sigma^{(0)}$, and $Y^{(0)}$.

while $t < T$ **do**

step 1: Sample $\left(\widetilde{Y}, (\alpha^{(a)})^2\right)$ from $p(\widetilde{Y}, \alpha^2 \mid \beta^{(t)}, \Phi^{(t)}, \Sigma^{(t)})$ in steps:

(a) Draw $(\alpha^{(a)})^2$ from its conditional prior $p(\alpha^2 \mid \Phi^{(t)}, \Sigma^{(t)})$. ▷ *Appendix C1*

(b) Draw $\widetilde{Y}$ from $p(\widetilde{Y} \mid (\alpha^{(a)})^2, \beta^{(t)}, \Phi^{(t)}, \Sigma^{(t)})$:

for $i := 1, \ldots, N$ **do**

for $k := 1, \ldots, K$ **do**

Sample Y_{ik} from $p(Y_{ik} \mid Y_{i,-k}, \beta^{(t)}, \Phi^{(t)}, \Sigma^{(t)})$.

end for

Set $\widetilde{Y}_i = \alpha^{(a)} Y_i$.

end for

step 2: Sample $\left(\beta^{(t+1)}, (\alpha^{(b)})^2\right)$ from $p(\beta, \alpha^2 \mid \widetilde{Y}, \Phi^{(t)}, \Sigma^{(t)})$ in steps:

(a) Draw $(\alpha^{(b)})^2$ from $p(\alpha^2 \mid \widetilde{Y}, \Phi^{(t)}, \Sigma^{(t)})$. ▷ *Appendix C2*

(b) Draw $\widetilde{\beta}$ from $p(\widetilde{\beta} \mid (\alpha^{(b)})^2, \widetilde{Y}, \Phi^{(t)}, \Sigma^{(t)})$.

(c) Set $\beta^{(t+1)} = \widetilde{\beta}/\alpha^{(b)}$.

step 3: Sample $(\widetilde{\eta}, \widetilde{u}_0)$ from $p(\widetilde{\eta}, \widetilde{u}_0 \mid \widetilde{Y}, \beta^{(t+1)}, \Phi^{(t)}, \Sigma^{(t)}, (\alpha^{(b)})^2)$ in steps:

(a) Draw $\widetilde{\eta}$ from $p(\widetilde{\eta} \mid \widetilde{Y}, \beta^{(t+1)}, \Phi^{(t)}, \Sigma^{(t)}, (\alpha^{(b)})^2)$. ▷ *Appendix C3*

(b) Draw $\widetilde{u}_0$ from $p(\widetilde{u}_0 \mid \widetilde{\eta}, \widetilde{Y}, \beta^{(t+1)}, \Phi^{(t)}, \Sigma^{(t)}, (\alpha^{(b)})^2)$.

step 4: Sample $\left(\Phi^{(t+1)}, \Sigma^{(t+1)}, (\alpha^{(c)})^2\right)$ from $p(\Phi, \Sigma, \alpha^2 \mid \widetilde{Z}, \widetilde{\eta}, \widetilde{u}_0)$,

where $\widetilde{Z} = (\widetilde{Z}_1, \ldots, \widetilde{Z}_N)'$ with $\widetilde{Z}_i = \widetilde{Y}_i - \alpha^{(b)} X_i \beta^{(t+1)} - \Gamma\widetilde{\eta} + \widetilde{u}_0$, for $i = 1, \ldots, N$,

using the following accept-reject procedure:

repeat

(a) Sample $\widetilde{\Phi}$ from $p(\widetilde{\Phi} \mid \widetilde{\eta})$. ▷ *Appendix C4*

(b) Sample $\widetilde{\Sigma}$ from $p(\widetilde{\Sigma} \mid \widetilde{Z})$.

(c) Retrieve $\alpha^{(c)}$ from $\widetilde{\Phi}$ and $\widetilde{\Sigma}$, as defined in Table 2.

(d) Compute $Y_i = \left(\widetilde{Z}_i + \alpha^{(c)} X_i \beta^{(t+1)} + \frac{\alpha^{(c)}}{\alpha^{(b)}} \left(\Gamma\widetilde{\eta} - \widetilde{u}_0\right)\right) / \alpha^{(c)}$,

until the following condition is satisfied, for all $i = 1, \ldots, N$:

$$\begin{cases} \max\{Y_{i1}, \ldots, Y_{iK}\} < 0 & \text{if } D_i = 0, \\ \max\{0, Y_{i1}, \ldots, Y_{iK}\} = Y_{ik} & \text{if } D_i = k. \end{cases} \tag{31}$$

Set $\Phi^{(t+1)} = \widetilde{\Phi} / \left(\alpha^{(c)}\right)^2$, $\Sigma^{(t+1)} = \widetilde{\Sigma} / \left(\alpha^{(c)}\right)^2$, and $Y^{(t+1)} = Y$.

return $\beta^{(t+1)}$, $\Phi^{(t+1)}$, $\Sigma^{(t+1)}$, and $Y^{(t+1)}$.

$t \leftarrow t + 1$.

end while

without adverse consequences.

Jiao and van Dyk (2015) point out two errors in the original sampling scheme derived in Imai and van Dyk (2005a), and offer an appropriate correction that ensures the sampler provides the correct stationary distribution. Since our procedure is an extension of the multinomial probit model, we apply a similar approach: In step 4, a transformation is applied to produce the parameters and latent variables of the identified model. It needs to be done starting from $\widetilde{Z}_i = \widetilde{Y}_i - \alpha^{(b)} X_i \beta^{(t+1)} - \Gamma\widetilde{\eta} + \widetilde{u}_0$, to take into account the information contained in the working parameter and in the latent variables sampled in steps 2 and 3. This transformation, however, might change the ordering of the latent variables, implying different observed decisions D. Since the transformation is made using the working parameter α^2, this parameter should be sampled conditional on the observational rule not being violated. To do this, we use an accept-reject procedure in step 4 to produce draws from $\widetilde{\Phi}$ and $\widetilde{\Sigma}$ that verify the corresponding condition in Eq. (31). The working parameter α^2 is a function of these two matrices and is computed differently depending on the identification restriction used, see Table 2.

4 Monte Carlo study

We investigate the properties of our MCMC sampler through a Monte Carlo experiment. We generate synthetic data that are similar to the real data used in our empirical application (Section 5), in terms of number of education and occupation decision categories, observed covariates and number of observations.

4.1 Experimental setup

Data generation. We simulate data from the differenced model specified in Eq. (8), combined with the decision rule in Eq. (7). Following our previous example, we specify a total of six alternative choices resulting from two levels of schooling and three occupation groups ($K = 5$ in the differenced system). For illustration purposes, we specify a single explanatory variable X, which varies between alternatives and individuals. The corresponding model is:

$$Y_i = X_i\beta + \Gamma\eta_i + \omega_i,$$

with

$$\beta = -0.2, \qquad X_{ij} \overset{iid}{\sim} \mathcal{U}(-3;\, 3)\,, \qquad \sigma_k^2 = 0.25,$$

for $i = 1, \ldots, N, j = 1, \ldots, 5$ and $k = 0, \ldots, 5$. The specification of the idiosyncratic variances implies that $\Sigma = 0.25(I_5 + \iota_5\iota_5')$. The $P = 3$ latent factors are generated and

allocated to the latent utilities as follows:

$$\eta_i = \begin{pmatrix} \eta_{i1} \\ \eta_{i2} \\ \eta_{i3} \end{pmatrix} \sim \mathcal{N}(0; \Phi), \qquad \Phi = \begin{pmatrix} 0.5 & -0.3 & 0.2 \\ -0.3 & 0.5 & 0.1 \\ 0.2 & 0.1 & 0.5 \end{pmatrix}, \qquad \Gamma = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}.$$

This parametrization results in samples with a well-balanced number of observations across the different alternatives. We generate datasets of sizes $N = 1,000$ and $5,000$.

Identification. We study each of the four identification restrictions outlined in Section 2.3 separately, and compare them. We contrast our results to those produced by the standard multinomial probit model, implemented in the `MNP` package (Imai and van Dyk, 2005b)—using both identification restrictions those authors studied as well.

Prior specification. We want to study the robustness of the algorithm to various priors—we do this by comparing four different specifications.

Note, however, that results from the different *identification* schemes cannot be directly compared, even if the same prior parameters are used: As explained in Section 3.3, a given specification of prior parameters induces different prior distributions in each of the four identification strategies because of the different parameter transformations used to obtain the identified model. Therefore, our posterior results should be contrasted with care, as differences in the posterior results might be due to differences in the induced prior distribution, rather than to differences in performances of the identification strategies.

There is a duality between the priors of the latent factors and the idiosyncratic error terms: Any prior distribution that implies a larger variance of the error terms in the expanded model will automatically result in a smaller relative impact of the latent factors in the identified model, because of the rescaling used to produce the parameters of the identified model from the expanded model. Therefore, it is important to study the role of the prior parameters of the unobserved components of the model jointly.

The four parametrizations we use are presented in Table 3. They use the same values for the parameters of the regression coefficients and the idiosyncratic variances, only varying in the parametrization of the distribution of Φ, the covariance matrix of the latent factors in the expanded model. Particularly, setting the number of degrees of freedom to $\nu_0 = P + 1 = 4$ implies that the marginal distributions of the correlations between the factors are uniform in the expanded model (Barnard et al., 2000). Increasing ν_0 to 6 pulls the correlations away from ± 1, allowing to prevent extreme cases with highly-correlated latent factors. Increasing the scale of the covariance matrix S_0 gives more weight to the latent factors compared to the error terms, everything else held equal.

For the benchmark standard MNP model, we specify very similar prior parameters: $B_0 = 10$, $\nu_0 = 6$, and $S_0 = I_3$.

Table 3: Prior specification in Monte Carlo experiment

	B_0	a_0	b_0	ν_0	S_0	t_0
spec 1	10	2	1	6	$4 \cdot I_3$	1
spec 2	10	2	1	4	I_3	1
spec 3	10	2	1	6	I_3	1
spec 4	10	2	1	4	$4 \cdot I_3$	1

Monte Carlo setup and MCMC tuning. For each identification strategy and each parametrization of the prior, we replicate the experiment 100 times, using different seeds of the random number generator to generate different data sets. In every replication, the model is sampled using 60,000 MCMC iterations, keeping only the last 50,000 iterations for posterior inference.

4.2 Simulation results

Separating latent factors from noise. The first criterion we use to assess our sampler is its capacity to correctly decompose the overall covariance matrix of the model into latent factors and noise. The share of unobserved heterogeneity captured by the latent factors can be measured by the following parameter:

$$\rho_j = \frac{\Gamma_j' \Phi \Gamma_j}{\Gamma_j' \Phi \Gamma_j + \sigma_0^2 + \sigma_j^2}, \tag{32}$$

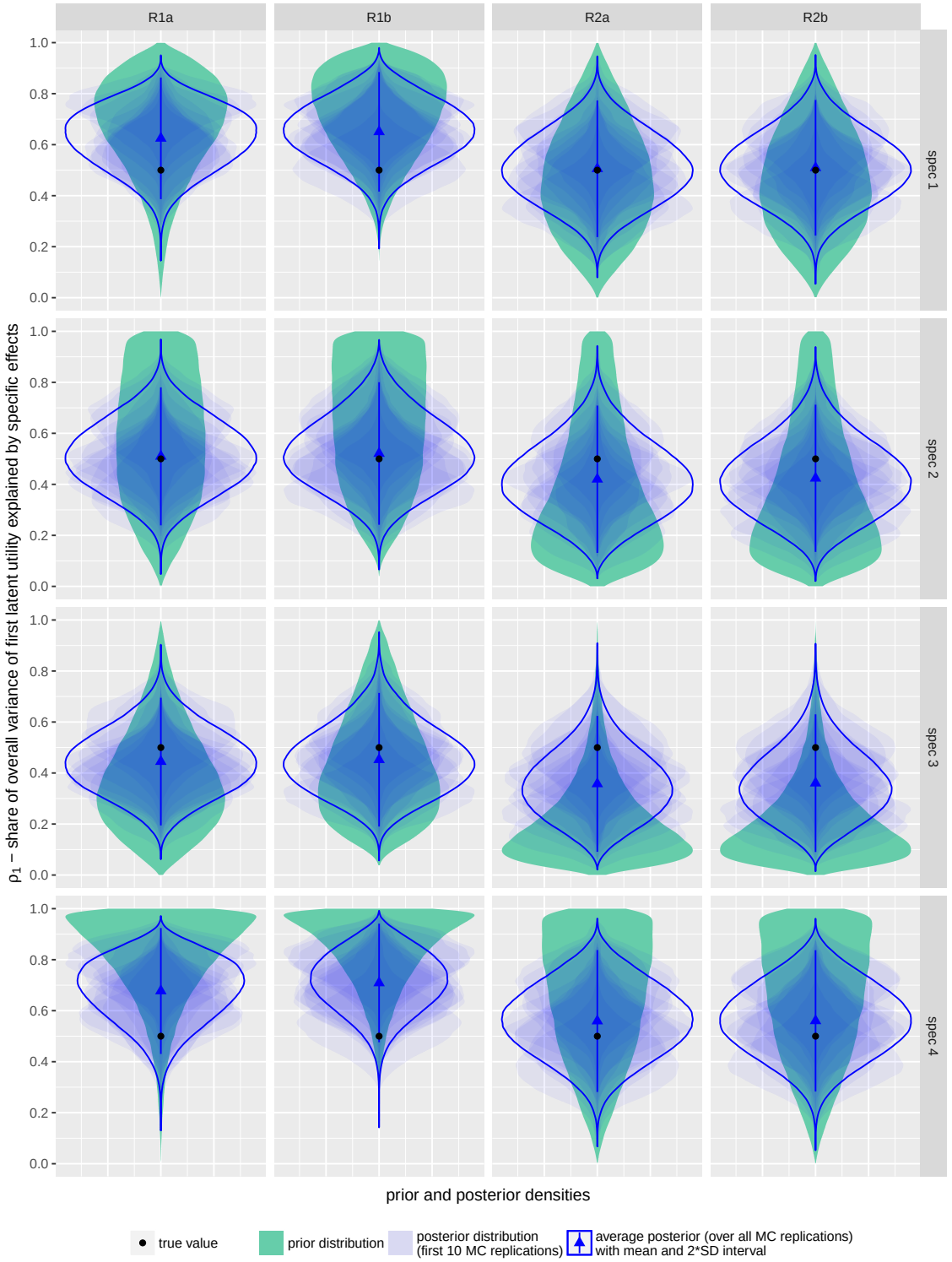
for each latent utility $j = 1, \ldots, K$, where Γ_j is the column vector of length P containing the j^{th} row of Γ.

Figs. 1 and 2 show the distribution of the first parameter ρ_1, both *a priori* and *a posteriori*, across the 100 Monte Carlo replications of the experiment, along with its true value in our synthetic data, for $N = 1,000$ and $5,000$ respectively. The vertical axes of these graphs show the values of ρ_1, which is distributed on the $[0, 1]$–interval, while the corresponding prior and posterior densities are shown on the horizontal axes. These violin plots display the densities as mirrored images, to emphasize where the mass of the distribution is located.

Each cell of these figures corresponds to a unique prior distribution, as prior parameters play a different role in the four identification strategies. This appears clearly here, when looking at the solid green shapes of the prior distributions across columns. For example, in the first row, the *same* specification (spec 1) implies a larger share of variance explained by the latent factors a priori for R1a and R1b than for R2a and R2b.

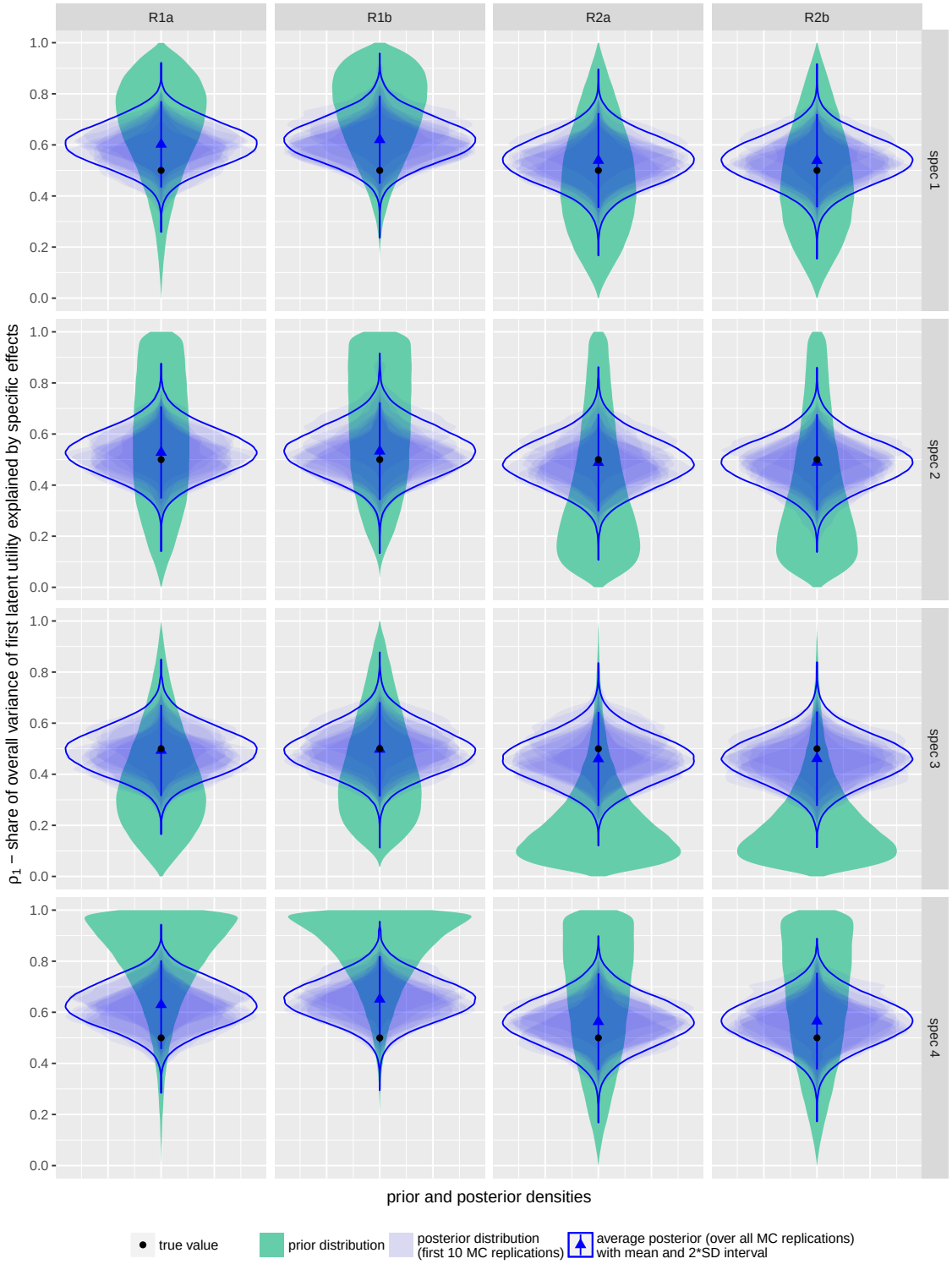
A posteriori, the sampler manages to learn from the data the share of overall variance that can be attributed to the latent factors, in the sense that the posterior distribution of ρ_1 (blue outline) deviates from the prior distribution in all cases, and the corresponding posterior mean (blue triangle) tends to the true value (black dot). There are, however, large disparities across prior specifications and identification restrictions. Extreme priors

Figure 1: Prior vs. posterior distribution of ρ_1, the share of overall variance of the first latent utility explained by the latent factors, for different prior specifications and identification restrictions. Model with 1000 observations.



Notes: Violin plots constructed as mirrored densities of the corresponding distributions. The four specifications of the prior parameters are shown in Table 3. Simulation of the prior based on 10^5 random draws. The prior and the average posterior have the same scale. The posteriors corresponding to the first 10 Monte Carlo replications (light blue) have been rescaled to fit in the graph area.

Figure 2: Prior vs. posterior distribution of ρ_1, the share of overall variance of the first latent utility explained by the latent factors, for different prior specifications and identification restrictions. Model with 5000 observations.



Notes: Violin plots constructed as mirrored densities of the corresponding distributions. The four specifications of the prior parameters are shown in Table 3. Simulation of the prior based on 10^5 random draws. The prior and the average posterior have the same scale. The posteriors corresponding to the first 10 Monte Carlo replications (light blue) have been rescaled to fit in the graph area.

that assume a lot of noise (e.g., spec 3 in R2*), or on the contrary very little noise (e.g., spec 4 in R1*) are difficult to escape from, and the sampler over- or under-estimates the share of noise in these cases. The impact of the prior diminishes when the number of observations increases. In estimations with $N = 5,000$, in Fig. 2, all models that use an identification restriction on the overall covariance matrix Ω (i.e., R2*) perform well. The posterior average is very close to the true noise share—even when the prior puts a large weight on very high or low noise shares. The restrictions on the idiosyncratic variances Σ (i.e., R1*) perform slightly worse in that case, as the effect of the prior is not entirely wiped out (see, for example, spec 1 or 4).

Therefore, a first general recommendation can be formulated: When the researcher does not have a strong prior for the share of noise relative to the latent factors, choosing the version of the sampler with an identification restriction on the overall covariance matrix Ω, such as R2a or R2b, is a safe option.

Recovering the latent structure of the model. The second criterion we use to assess our sampler's performance is the root mean squared error (rMSE) of the inferred parameters. The rMSEs for all four prior specifications across the four identification restrictions are shown in Table 4. As a benchmark, we also provide the results obtained from the standard MNP model (with the existing `MNP` package, Imai and van Dyk, 2005b) on the same artificial data sets, with both the identification restriction on the first element of Ω (MNPa) and on its trace (MNPb). For a fair comparison of the methods, the identification strategies should be compared pairwise depending on whether the restriction is imposed on a single element (MNPa/R1a/R2a) or on a combination of the diagonal elements ("trace restriction", MNPb/R1b/R2b).

For the regression coefficient β, we simply report the average rMSE over the 100 Monte Carlo replications. For the remaining parameters, Table 4 shows an aggregate measure corresponding to the average rMSE within each group of parameters Σ, Φ and Ω.[12] The benchmark MNP model does not allow to decompose the overall covariance matrix, therefore only results for β and Ω are provided.

The results show that the MCMC sampler generally retrieves the true coefficient β in a very satisfactory way, as the rMSE is small in all cases. For all other parameters, important differences can be observed across identification restrictions. The trace restriction provides better results for β than the element restriction in all cases, thus confirming that the results of Burgette and Nordheim (2012) for the standard MNP model also apply to our approach. This is also noteworthy for the Σ-restrictions of type R1*, where the trace restriction (R1b) provides a large improvement over the element restriction (R1a)—there, the highest rMSEs are observed. For Σ and especially Φ, placing the restriction on the idiosyncratic variances Σ is associated with more error than the restriction on the full covariance matrix Ω. The identification restrictions using the full covariance matrix Ω produce much lower rMSEs,

12. This aggregate measure is only used to limit the number of parameters to be displayed (e.g., 15 unique parameters in Ω only), and does not alter the main conclusions of the analysis.

Table 4: Comparing root mean squared errors of model parameters across identification restrictions and prior specifications in the Monte Carlo study.

(a) Number of observations: 1,000

		MNPa	**MNPb**	**R1a**	**R1b**	**R2a**	**R2b**
β	spec 1	0.032	0.015	0.073	0.046	0.026	0.013
	spec 2	0.032	0.015	0.048	0.026	0.027	0.013
	spec 3	0.034	0.016	0.037	0.024	0.026	0.014
	spec 4	0.035	0.015	0.115	0.073	0.026	0.013
Σ	spec 1			0.301	0.134	0.128	0.121
	spec 2			0.237	0.121	0.145	0.132
	spec 3			0.184	0.112	0.162	0.145
	spec 4			0.404	0.144	0.129	0.125
Φ	spec 1			0.506	0.418	0.145	0.173
	spec 2			0.337	0.255	0.154	0.176
	spec 3			0.234	0.205	0.172	0.184
	spec 4			0.940	0.671	0.154	0.182
Ω	spec 1	0.431	0.301	0.959	0.683	0.326	0.223
	spec 2	0.424	0.302	0.583	0.393	0.337	0.236
	spec 3	0.448	0.303	0.383	0.325	0.314	0.239
	spec 4	0.456	0.303	1.804	1.125	0.358	0.225

(b) Number of observations: 5,000

		MNPa	**MNPb**	**R1a**	**R1b**	**R2a**	**R2b**
β	spec 1	0.015	0.007	0.059	0.030	0.013	0.006
	spec 2	0.015	0.007	0.035	0.017	0.013	0.006
	spec 3	0.015	0.007	0.028	0.015	0.014	0.006
	spec 4	0.015	0.007	0.087	0.040	0.014	0.006
Σ	spec 1			0.196	0.094	0.083	0.086
	spec 2			0.143	0.086	0.087	0.090
	spec 3			0.119	0.084	0.089	0.093
	spec 4			0.274	0.099	0.087	0.091
Φ	spec 1			0.359	0.235	0.094	0.106
	spec 2			0.220	0.155	0.092	0.103
	spec 3			0.162	0.133	0.096	0.105
	spec 4			0.574	0.313	0.103	0.112
Ω	spec 1	0.215	0.150	0.718	0.392	0.188	0.131
	spec 2	0.213	0.149	0.402	0.241	0.195	0.132
	spec 3	0.218	0.149	0.303	0.207	0.190	0.130
	spec 4	0.218	0.149	1.167	0.518	0.193	0.131

Notes: The rMSE presented here is the root mean squared error across the 100 Monte Carlo replications. For β, we simply show the rMSE. For the other parameters, Σ, Φ, and Ω, where there are several parameters of the same type (e.g., the 6 elements of Σ), we report the average rMSE within this group. The columns correspond to different identification restrictions. "MNPa" and "MNPb" correspond to the MNP package with an element and trace restriction, respectively. For the remaining 4 columns, see Table 1. Cells in grey show the smallest rMSE in each row of the table.

at an entirely different order of magnitude. Again, the restriction that combines multiple elements of the trace, R2b, works best for most parameters.

The general lesson of this simulation exercise is that even though all four restrictions yield a theoretically identified model, large disparities between their respective performances prevail. The restriction on the elements of the trace of Ω (R2b) is most stable in recovering the parameters (Table 4) and, more importantly, in retrieving the share of the unobserved variance driven by the specific latent factors rather than noise (Fig. 2).

5 Empirical application

To illustrate the empirical relevance and the applicability of our method, we study the determinants of educational and occupational choice in a subsample of the American Community Survey of 2015 (1% sample, Ruggles et al., 2010). The number of alternatives and sample size are similar to the setup of our Monte Carlo experiment.

5.1 Data description

In the American Community Survey, we focus on individuals aged 34-35 who are in the labor force. Among those who have valid entries for the outcome and explanatory variables, we select a random subsample of 1,000 individuals. The outcome variable is the joint choice of education and occupation. Education can be "Low ed" with up to 12 years of education (high school graduation), or "High ed" with some college or more. Occupations are aggregated into three groups on the basis of 2-digit SOC occupation codes: "Blue collar," "Services," and "Business/Management/STEM." Forming all education-occupation combinations yields the six alternatives listed in Table 5. The baseline in our analysis is taken as "Low ed – Blue collar." Recall that the latent factors in η_i correspond to differences of the occupation- or education-specific taste shocks with respect to their base-level: here, η_1 corresponds to the difference between the taste for "Services" vs. "Blue collar", η_2 to "Business" vs. "Blue collar", and η_3 to the difference between the specific taste for "High ed" vs. "Low ed."

As choice-specific covariates, we use the local labor market characteristics of the occupation-education pairs—determined among workers that are older than our target group, but who live in the same area. Within each state, the Census Bureau defines PUMAs (Public Use Microdata Areas), an area of contiguous territory containing 100,000+ residents. Within each PUMA, we generate averages, for workers aged 36-65, of the following two characteristics: wage income (`incwage`) and the share of workers of the same gender as the respondent in this occupation-education pair (`gendershare`). In addition to these alternative-specific covariates, we use indicators for the respondent's person-fixed characteristics of marital status (`married`) and race (`non-white`). Non-white is defined as self-classification as either African-American, Hispanic, or Asian.

Table 5: Available Choices in ACS Data

Choice		#	Freq	Percent	η_1	η_2	η_3
Low ed	Blue collar	0	160	16	.	.	.
Low ed	Services	1	150	15	1	0	.
Low ed	Bus/Managmt/STEM	2	32	3.2	0	1	.
High ed	Blue collar	3	74	7.4	.	.	1
High ed	Services	4	383	38.3	1	0	1
High ed	Bus/Managmt/STEM	5	201	20.1	0	1	1
Total		.	1000	100	.	.	.

Note: The chosen baseline in our analysis is "Low ed – Blue Collar." Recall that η_1 and η_2 are differences in tastes for occupations, with baseline "Blue collar" (indicated with "."), so η_1 is the difference between the taste for "Services" vs. "Blue collar", η_2 between "Business" and "Blue collar," and η_3 is the difference in education taste between "High ed" and "Low ed."

5.2 Results

The results of the application are shown in Table 6. The algorithm was implemented with the identification restriction R2b, using 100,000 MCMC iterations (after a burn-in period of 10,000 iterations). The prior corresponds to "spec 4" in Table 3. We tried alternative prior specifications. In line with the results from the Monte Carlo study, we find that the choice of the prior does not influence the results in any noticeable way as long as the prior is not too informative about the share of noise in the latent part of the model.[13]

For all parameters, we report the mean and standard deviation of the posterior distribution, as well as the lower and upper endpoints of the 95% highest posterior density (HDP) interval interval.[14] The coefficients reveal that conditional on the other covariates, choices 2 and 3 are rather infrequent—see the negative 2nd and 3rd intercepts. It is generally unusual to select into business/management/STEM jobs together with low education, and (to a lesser extent) also a blue collar occupation with high education. Marital status has only a weak or noisy influence on choices (the HDP intervals include zero). But race significantly impacts the choice of occupation-education pairs, as do the alternatives' average income among older workers (`incwage`), and the share of the same gender (`gendershare`). The signs of the coefficients are in line with our expectations: workers are more likely to choose an occupation that has higher wages (among the other workers in their area), and which has a larger share of their own gender as well.

We can now turn to interpret the covariance matrix of the unobserved heterogeneity. Recall that the σ_k^2 elements correspond to the variance of the noise that remains after accounting for the structure of the latent factors. Comparing the magnitude of the σ_k^2 elements to the variance of the factors in Φ shows that a substantial amount of the unobserved variation stems from the factors rather than the error terms in our example. To see this more clearly, we can use the means of the posterior distribution to compute ρ in

13. Additional results available upon request.

14. The shortest interval that covers 95% of the empirical cumulative distribution function of the parameters' sample values, comparable to a confidence interval.

Table 6: Empirical results on ACS Data

	Mean	SD	[95%	HPD]
intercept_1	-0.16	0.19	-0.53	0.22
intercept_2	-0.87	0.30	-1.47	-0.30
intercept_3	-0.35	0.18	-0.72	-0.00
intercept_4	0.51	0.13	0.26	0.78
intercept_5	-0.22	0.20	-0.63	0.17
married_1	-0.18	0.11	-0.41	0.05
married_2	0.05	0.13	-0.20	0.30
married_3	0.02	0.15	-0.27	0.31
married_4	-0.12	0.16	-0.45	0.20
married_5	-0.02	0.13	-0.26	0.24
non-white_1	-0.44	0.14	-0.71	-0.16
non-white_2	0.07	0.09	-0.10	0.24
non-white_3	-0.51	0.10	-0.71	-0.31
non-white_4	0.28	0.12	0.06	0.52
non-white_5	-0.37	0.13	-0.64	-0.13
incwage	0.26	0.11	0.04	0.47
gendershare	0.99	0.15	0.70	1.26
σ_0^2	0.15	0.10	0.02	0.33
σ_1^2	0.36	0.27	0.01	0.87
σ_2^2	0.21	0.17	0.01	0.57
σ_3^2	0.16	0.12	0.01	0.41
σ_4^2	0.09	0.05	0.01	0.19
σ_5^2	0.13	0.10	0.01	0.34
$\Phi_{1:1}$	0.45	0.19	0.11	0.82
$\Phi_{1:2}$	0.16	0.18	-0.16	0.53
$\Phi_{1:3}$	-0.39	0.21	-0.77	0.00
$\Phi_{2:2}$	0.48	0.24	0.08	0.93
$\Phi_{2:3}$	-0.21	0.26	-0.76	0.32
$\Phi_{3:3}$	0.67	0.29	0.14	1.24
$\Phi_{1:2}^{(\text{cor})}$	0.33	0.35	-0.37	0.90
$\Phi_{1:3}^{(\text{cor})}$	-0.67	0.21	-0.96	-0.24
$\Phi_{2:3}^{(\text{cor})}$	-0.36	0.41	-0.98	0.47

Note: Mean and standard deviation of the posterior distribution, with identification restriction R2b and prior $B_0 = 10, a_0 = 2, b_0 = 1, \nu_0 = 4, S_0 = 4 \cdot I_3, t_0 = 1$. We report lower and upper limits of the 95% highest posterior density (akin to a confidence interval). See Table 5 for the definition of alternatives. Elements of Φ correspond to the covariance between the η (see also Table 5).

Eq. (32). The implied shares of variance explained by the latent factors, for each of the latent utilities, are $\rho_1 = 48\%, \rho_2 = 59\%, \rho_3 = 70\%, \rho_4 = 60\%, \rho_5 = 72\%$. These shares are relatively large and demonstrate the importance of occupation or education-specific factors that drive the choice. The data hold information that can be backed out rather than only calling all terms in Ω "noise."

Throughout this text, we have motivated the need for occupation- or education-specific latent factors with the example of unobserved taste shocks that apply to specific occupations or education levels. Along this example, the structure for these latent factors arose from the fact that each occupation had an unobserved person-specific component that appeared in all alternatives that contained this occupation. We will continue this example in the interpretation below. For the sake of completeness, however, we note that the latent factors could reflect not only taste differences, but other unobserved characteristics that follow the same structure. They could, for example, also reflect unobserved occupation- or education-specific ability differences, depending on the data used. What is important for our interpretation is that these *choice-specific* unobserved tastes or abilities are picked up by the structure, which in turn lets us relate them *to each other* and *to the remaining noise*. The economic content in the decision-making process is given by these recurrent factors that drive choices in a traceable way. In our example, we conclude that unobserved tastes or characteristics for specific occupations and education shape 50-70% of the variation in observed choices.

Within Φ, the variation in latent preferences seems to come relatively evenly from preferences for occupations and schooling, with a slightly larger role for schooling (see the variance terms $\Phi_{1:1}$ to $\Phi_{3:3}$). More interestingly, we can examine the off-diagonal terms of Φ to learn about the relationship between the different factors. For convenience, we report the correlation of the factors as $\Phi^{(\text{cor})}$ in Table 6. Begin with $\Phi^{(\text{cor})}_{1:2}$. Because this parameter's HDP interval includes zero, we conclude that there is no evidence in the data that η_1 and η_2 are correlated. This implies that the *relative* preferences for "Services" and "Business" are not correlated: if a worker prefers service jobs relative to blue collar jobs (large η_1), we cannot infer that he also prefers business/management jobs to blue collar jobs (we cannot conclude that η_2 will also be large). In larger models with more occupation categories, this type of analysis could reveal which occupations are seen as closer substitutes than others, while the preference for education is controlled for.

Next, observe that $\Phi^{(\text{cor})}_{1:3}$ is negative. Recall that η_3 corresponds to the relative utilities of the two education levels. The negative correlation of η_1 and η_3 shows us that individuals who prefer service sector over blue collar jobs (large η_1) *dislike* the higher education level relative to the lower one (low η_3). Alternatively phrased, individuals with a strong taste for higher education do not prefer service sector over blue collar jobs. Would individuals with a taste for high education then prefer management-type jobs over blue-collar jobs? Again, we do not find evidence for this in the data, once we hold alternative-specific and personal characteristics constant: the posterior of $\Phi^{(\text{cor})}_{2:3}$ contains zero in its 95% HPD interval. This means that individuals who prefer high education do *not* have a stronger

preference for management jobs over blue collar jobs, relative to those who prefer low education.

These interpretations of the covariance matrix Φ are informative for the decision-making process, as they go beyond an observation of how frequently choices are made together (cf. Table 5). We have conditioned the decision on both alternative-specific and person-specific covariates, and within this model learned about the correlation between taste preferences for the occupation options and education levels separately.

6 Conclusion

This paper develops a multinomial probit model with latent factors that can flexibly be allocated to different alternatives. The factor allocation can directly reflect an economic decision structure, if researchers have *a priori* knowledge of such a structure. A prime example of this setting, which we investigate empirically, is for choices that are taken jointly, so that alternatives reflect pairs of choices.

Our contribution focuses on the researcher's ability to interpret the MNP results in light of the underlying economic decision process, without requiring data for a factor measurement system. We achieve this through three steps: The first is that factors are allocated to alternatives in a way that can directly reflect the economic model. Secondly, the resulting parametrization of MNP yields a manageable covariance matrix for both estimation and interpretation. The usual problem of the quickly increasing number of parameters is addressed through the factor structure, and the estimated covariance matrix of the factors provides researchers with information about how different latent decision-drivers are correlated. Finally, we fully guarantee economic interpretability with our theoretical proofs and discussion of empirical identification.

Our simulation exercise shows that our approach manages to decompose the unobserved heterogeneity into noise and economic content (latent factors). How well it does so depends on the identification restriction used. The recommendation for practitioners is to favor the "trace" restriction on the full covariance matrix (R2b). Other than the variance decomposition, the sampler also retrieves the remaining parameters (coefficients β and overall covariance matrix Ω) well. It does so with a significantly lower rMSE than the state-of-the-art non-structured MNP model.

The fact that our proposed algorithm outperforms an uninformative MNP in a setting where the economic model represents a simple structure shows that our sampler should be prioritized whenever this type of structure is available. Our approach addresses the computational and interpretation challenges that remain with existing MNP approaches, and offers an attractive alternative that provides economically meaningful results without additional data requirements, while being computationally at least as efficient as existing MNP implementations.

Acknowledgments

This paper was previously circulated under the title "A Parsimonious Multinomial Probit Model for the Study of Joint Decisions". It was presented at the 11th World Congress of the Econometric Society in Montreal, at the Research Seminar of the Institute for Statistics and Mathematics at the Vienna University of Economics and Business (WU, Austria), at the Econometric Conference in Honor of François Laisney (Strasbourg, France), at the Seminar of the Centre for Applied Microeconometrics (CAM, University of Copenhagen), and discussed in the Education Group at the Department of Economics, University of Copenhagen. The authors are very grateful for all the comments received at these conferences and seminars, which substantially helped improve the methodology.

This research was funded by the Danish Council for Independent Research and the Marie Curie programme COFUND under the European Union's Seventh Framework Programme for research, technological development and demonstration. Grant-ID DFF–4091-00246 for Rémi Piatek. Grant-ID DFF–4091-00240 for Miriam Gensowski.

References

Anderson, T. W., and H. Rubin. 1956. "Statistical Inference in Factor Analysis." Chap. 3 in *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability,* edited by J. Neyman, 5:111–150. Berkeley: University of California Press.

Barnard, J., R. E. McCulloch, and X.-L. Meng. 2000. "Modeling Covariance Matrices in Terms of Standard Deviations and Correlations, with Application to Shrinkage." *Statistica Sinica* 10:1281–1311.

Ben-Akiva, M., J. Walker, A. T. Bernardino, D. A. Gopinath, T. Morikawa, and A. Polydoropoulou. 2002. "Integration of Choice and Latent Variable Models." In *In Perpetual Motion: Travel Behavior Research Opportunities and Application Challenges,* 431–470. Amsterdam: Pergamon.

Bhat, C. R., and S. K. Dubey. 2014. "A New Estimation Approach to Integrate Latent Psychological Constructs in Choice Modeling." *Transportation Research Part B* 67:68–85. doi:10.1016/j.trb.2014.04.011.

Bolduc, D. 1992. "Generalized Autoregressive Errors in the Multinomial Probit Model." *Transportation Research Part B* 26 (2): 155–170. doi:10.1016/0191-2615(92)90005-H.

Börsch-Supan, A., V. Haijvassiliou, L. J. Kotlikoff, and J. N. Morris. 1992. "Health, Children, and Elderly Living Arrangements: A Multiperiod-Multinomial Probit Model with Unobserved Heterogeneity and Autocorrelated Errors." In *Topics in the Economics of Aging,* edited by D. A. Wise, 79–108. doi:10.3386/w3343.

Bunch, D. S. 1991. "Estimability in the Multinomial Probit Model." *Transportation Research Part B: Methodological* 25 (1): 1–12. doi:10.1016/0191-2615(91)90009-8.

Burgette, L. F., and E. V. Nordheim. 2012. "The Trace Restriction: An Alternative Identification Strategy for the Bayesian Multinomial Probit Model." *Journal of Business & Economic Statistics* 30 (3): 404–410. doi:10.1080/07350015.2012.680416.

Cripps, E., D. G. Fiebig, and R. Kohn. 2009. "Parsimonious Estimation of the Covariance Matrix in Multinomial Probit Models." *Econometric Reviews* 29 (2): 146–157. doi:10.1080/07474930903382158.

Dansie, B. R. 1985. "Parameter Estimability in the Multinomial Probit Model." *Transportation Research Part B: Methodological* 19 (6): 526–528. doi:10.1016/0191-2615(85)90047-5.

Daziano, R. A., and D. Bolduc. 2013. "Incorporating Pro-environmental Preferences towards Green Automobile Technologies through a Bayesian Hybrid Choice Model." *Transportmetrica A: Transport Science* 9 (1): 74–106. doi:10.1080/18128602.2010.524173.

Elrod, T., and M. P. Keane. 1995. "A Factor-Analytic Probit Model for Representing the Market Structure in Panel Data." *Journal of Marketing Research* 32 (1): 1–16. doi:10.2307/3152106.

Fu, X., and Z. Juan. 2017. "Estimation of Multinomial Probit-Kernel Integrated Choice and Latent Variable Model: Comparison on One Sequential and Two Simultaneous Approaches." *Transportation* 44:91–116. doi:10.1007/s11116-015-9626-x.

Geweke, J., M. Keane, and D. Runkle. 1994. "Alternative Computational Approaches to Inference in the Multinomial Probit Model." *The Review of Economics and Statistics* 76 (4): 609–632. doi:10.2307/2109766.

Imai, K., and D. A. van Dyk. 2005a. "A Bayesian Analysis of the Multinomial Probit Model using Marginal Data Augmentation." *Journal of Econometrics* 124 (2): 311–334. doi:10.1016/j.jeconom.2004.02.002.

———. 2005b. "MNP: R Package for Fitting the Multinomial Probit Model." *Journal of Statistical Software* 14 (3): 1–32.

Jiao, X., and D. A. van Dyk. 2015. "A Corrected and More Efficient Suite of MCMC Samplers for the Multinomal Probit Model." *Working Paper:* 1–20. arXiv: 1504.07823.

Keane, M. P. 1992. "A Note on Identification in the Multinomial Probit Model." *Journal of Business & Economic Statistics* 10 (2): 193–200. doi:10.1080/07350015.1992.10509898.

Keane, M. P., and K. I. Wolpin. 1997. "The Career Decisions of Young Men." *Journal of Political Economy* 105 (3): 473–522. doi:10.1086/262080.

Lee, D. 2005. "An Estimable Dynamic General Equilibrium Model of Work, Schooling, and Occupational Choice." *International Economic Review* 46 (1): 1–34. doi:10.1111/j.0020-6598.2005.00308.x.

Liu, C., D. B. Rubin, and Y. N. Wu. 1998. "Parameter Expansion to Accelerate EM : The PX-EM Algorithm." *Biometrika* 85 (4): 755–770. doi:10.1093/biomet/85.4.755.

Liu, J. S., and Y. N. Wu. 1999. "Parameter Expansion for Data Augmentation." *Journal of the American Statistical Association* 94 (448): 1264–1274. doi:10.1080/01621459.1999.10473879.

Magnus, J. R., and H. Neudecker. 2007. *Matrix Differential Calculus with Applications in Statistics and Econometrics.* Third Edit. John Wiley & Sons Ltd.

McCulloch, R. E., N. G. Polson, and P. E. Rossi. 2000. "A Bayesian Analysis of the Multinomial Probit Model with Fully Identified Parameters." *Journal of Econometrics* 99 (1): 173–193. doi:10.1016/S0304-4076(00)00034-8.

McCulloch, R. E., and P. E. Rossi. 1994. "An Exact Likelihood Analysis of the Multinomial Probit Model." *Journal of Econometrics* 64 (1-2): 207–240. doi:10.1016/0304-4076(94)90064-7.

———. 2000. "Reply to Nobile." *Journal of Econometrics* 99:347–348. doi:10.1016/S0304-4076(00)00036-1.

Meng, X.-L., and D. A. van Dyk. 1999. "Seeking Efficient Data Augmentation Schemes via Conditional and Marginal Augmentation." *Biometrika* 86 (2): 301–320. doi:10.1093/biomet/86.2.301.

Nobile, A. 1998. "A Hybrid Markov Chain for the Bayesian Analysis of the Multinomial Probit Model." *Statistics and Computing* 8 (3): 229–242. doi:10.1023/A:1008905311214.

———. 2000. "Comment: Bayesian Multinomial Probit Models with a Normalization Constraint." *Journal of Econometrics* 99:335–345. doi:10.1016/S0304-4076(00)00035-X.

Nobile, A., C. R. Bhat, and E. I. Pas. 1997. "A Random Effects Multinomial Probit Model of Car Ownership Choice." *Case Studies in Bayesian Statistics,* Lecture Notes in Statistics, III:419–434. doi:10.1007/978-1-4612-2290-3_13.

Ruggles, S., J. T. Alexander, K. Genadek, R. Goeken, M. B. Schroeder, and M. Sobek. 2010. *Integrated Public Use Microdata Series: Version 5.0 [Machine-readable database].* Minneapolis: University of Minnesota.

Tanner, M. A., and W. H. Wong. 1987. "The Calculation of Posterior Distributions by Data Augmentation." *Journal of the American Statistical Association* 82 (398): 528–540. doi:10.1080/01621459.1987.10478458.

Train, K. E. 2009. *Discrete Choice Methods with Simulation.* Second Edi. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511805271.

Van Dyk, D. A. 2010. "Marginal Markov Chain Monte Carlo Methods." *Statistica Sinica* 20 (4): 1423–1454.

Van Dyk, D. A., X.-L. Meng, and C. Liu. 2001. "The Art of Data Augmentation." *Journal of Computational and Graphical Statistics* 10 (1): 1–50. doi:10.1198/10618600152418584.

Van Dyk, D. A., and T. Park. 2008. "Partially Collapsed Gibbs Samplers: Theory and Methods." *Journal of the American Statistical Association* 103 (482): 790–796. doi:10.1198/016214508000000409.

———. 2009. "Partially Collapsed Gibbs Samplers: Illustrations and Applications." *Journal of Computational and Graphical Statistics* 18 (2): 283–305. doi:10.1198/jcgs.2009.08108.

Yai, T., S. Iwakura, and S. Morichi. 1997. "Multinomial Probit with Structured Covariance for Route Choice Behavior." *Transportation Research Part B: Methodological* 31 (3): 195–207. doi:10.1016/S0191-2615(96)00025-2.

A Proofs

A1 Proof of Proposition 2.1

Proof. Since G is made of P linearly independent rows of $\Delta_K \Gamma^\star$, it is full rank with $\text{rank}\,(G) = P$. Therefore, GG' is nonsingular and the Moore-Penrose pseudoinverse of G can be constructed as $H = G'(GG')^{-1}$. This generalized inverse is a *right inverse*, such that $GH = I_P$, and $HG = G'(GG')^{-1}G = P_{G'}$ is the projection matrix on the span of the columns of G'. As a consequence, the transformation in Eq. (11) can be expressed as $\Gamma\eta_i = \Delta_K \Gamma^\star HG\,\eta_i^\star = [P_{G'}(\Delta_K\Gamma^\star)']'\eta_i^\star = \Delta_K\Gamma^\star\eta_i^\star$, where the last equality comes from the fact that G is made of $P = \text{rank}\,(\Delta_K\Gamma^\star)$ linearly independent rows of $\Delta_K\Gamma^\star$, therefore $P_{G'}(\Delta_K\Gamma^\star)' = (\Delta_K\Gamma^\star)'$. This result is also valid if G is made of P linearly independent rows that are linear combinations of rows of $\Delta_K\Gamma^\star$. □

A2 Proof of Proposition 2.2

The proof proceeds in two steps: First, it is shown that the covariance matrix Φ and the idiosyncratic variance of the baseline level σ_0^2 are identified from the lower triangular elements of Ω—excluding the diagonal elements. With the identification of these parameters

in hand, the remaining idiosyncratic variances $\sigma_1^2, \ldots, \sigma_K^2$ are identified from the diagonal elements of Ω in a second step.

The model is identified if and only if the system of linear equations corresponding to $\Omega = \Gamma \Phi \Gamma' + \Sigma$ is full rank. The overall covariance matrix Ω can be vectorized as:[15]

$$\begin{aligned} \operatorname{vec}(\Omega) &= \operatorname{vec}\left(\Gamma \Phi \Gamma'\right) + \sigma_0^2 \, \iota_{K^2} + \operatorname{vec}\left(\operatorname{diag}(\sigma_1^2, \ldots, \sigma_K^2)\right), \\ &= \underbrace{\begin{pmatrix} \Gamma \otimes \Gamma & \iota_{K^2} \end{pmatrix}}_{S} \underbrace{\begin{pmatrix} e_1 \otimes e_1 & e_2 \otimes e_2 & \cdots & e_K \otimes e_K \end{pmatrix}}_{E} \begin{pmatrix} \operatorname{vec}(\Phi) \\ \sigma_0^2 \\ \vdots \\ \sigma_{K^2}^2 \end{pmatrix}, \end{aligned} \tag{A1}$$

where e_k is the K-vector containing only zeros besides its kth element that is equal to one, for $k = 1, \ldots, K$, and using the result that $\operatorname{vec}(\Gamma \Phi \Gamma') = (\Gamma \otimes \Gamma) \operatorname{vec}(\Phi)$, where $\otimes$ denotes the Kronecker product.

Identification is achieved if and only if the matrix $A = \begin{pmatrix} S & E \end{pmatrix}$ is full rank, or equivalently, if and only if the determinant of $A'A$ is different from zero. Using some results on the determinant of a partitioned matrix, it follows that

$$\left|A'A\right| = \left|E'E\right| \times \left|E'E - S'E(E'E)^{-1}E'A\right| = \left|(M_E S)'(M_E S)\right|, \tag{A2}$$

because $|E'E| = |I_K| = 1$, and $M_E = I_{K^2} - E(E'E)^{-1}E' = I_{K^2} - EE'$ is the projection matrix on the orthogonal of the subspace spanned by the columns of E.[16] The last determinant in Eq. (A2) is different from zero if and only if $M_E S$ is full rank. Given the structure of M_E, the projection $M_E S$ is equivalent to transforming the matrix E such that all the rows corresponding to the diagonal elements of the overall matrix Ω are set to zero. This result is used in Proposition A1 to show that $M_E S$ is full rank. Finally, Proposition A2 shows that the redundant parameters and equations in the system of linear equations Eq. (A1) are innocuous for the full rank condition of the system.

Proposition A1. *If Γ fulfills Assumption 2.1, then the matrix $M_E S = \begin{pmatrix} M_E(\Gamma \otimes \Gamma) & M_E \iota_{K^2} \end{pmatrix}$, obtained by projecting S on the orthogonal of the subspace spanned by the columns of E, is full rank.*

Proof. The Kronecker product $\Gamma \otimes \Gamma$ is made of K blocks $B_k \equiv \Gamma_{k\bullet} \otimes \Gamma$ stacked one on top of each other, where $\Gamma_{k\bullet}$ is the kth row of Γ, for $k = 1, \ldots, K$. Each block B_k is obtained by rescaling and repeating the columns of Γ, and/or adding zero columns. Given that every row of Γ is a linear combination of some of its other rows (see condition 2 of Assumption 2.1), the same applies to B_k. As a consequence, it is possible to delete one row within each block B_k without reducing the rank of the overall matrix $\Gamma \otimes \Gamma$.

15. The $\operatorname{vec}(\cdot)$ operator stacks the columns of the corresponding matrix, such that if $X = \begin{pmatrix} a & c \\ b & d \end{pmatrix}$, then $\operatorname{vec}(X) = \begin{pmatrix} a & b & c & d \end{pmatrix}'$, whereas the $\operatorname{vech}(\cdot)$ operator used later stacks the lower triangular elements of the corresponding matrix, such that $\operatorname{vech}(X) = \begin{pmatrix} a & b & d \end{pmatrix}'$.

16. The projection matrix M_E is symmetric and idempotent matrix, i.e., $M_E = M_E'$ and $M_E M_E = M_E$.

Projecting $\Gamma \otimes \Gamma$ on the orthogonal of the span of the columns of E is equivalent to nullifying the kth row of each block B_k, for $k = 1, \ldots, K$. Since $\Gamma \otimes \Gamma$ is full rank and has P^2 linearly independent rows,[17] the argument of the previous paragraph implies that $M_E(\Gamma \otimes \Gamma)$ is full rank.

The column vector $M_E \iota_{K^2} = \iota_{K^2} - \begin{pmatrix} e_1' & \cdots & e_K' \end{pmatrix}'$ is not a linear combination of the columns of $M_E(\Gamma \otimes \Gamma)$. This can be seen from the structure of Γ (see also Eq. (15)): because it contains the identity matrix, at least $P - 1$ of its columns would need to be added to produce the required 1s in $\iota_K - e_k$, but this sum would imply 2s in the remaining elements, thus making it impossible to create $\iota_K - e_k$ from a linear combination of the columns of $M_E(\Gamma \otimes \Gamma)$. It is enough to apply this argument to one block B_k to show that $M_E \iota_{K^2}$ is linearly independent of the columns of $M_E(\Gamma \otimes \Gamma)$.

Therefore, $M_E S$ has $P^2 + 1$ linearly independent columns, i.e., it is full rank. □

Proposition A2. *Omitting redundant parameters and redundant equations in Eq. (A1) does not affect the full rank of the resulting matrix.*

Proof. Because of the symmetry of the covariance matrices Ω and Φ, some elements are redundant and can be omitted. This is done using the vech $(\cdot)$ operator, which stacks the lower triangular elements of the corresponding matrix column-wise. The two operators are related through the duplication matrix D_n, of dimension $n^2 \times n(n+1)/2$, which transforms vech $(\cdot)$ into vec $(\cdot)$, such that for any symmetric matrix A, vec $(A) = D_n$ vech (A), see Magnus and Neudecker (2007, p. 56-57).

Equation (A1) can be re-expressed as:

$$\text{vech}\,(\Omega) = D_K^+ \underbrace{\begin{pmatrix} (\Gamma \otimes \Gamma) D_P & \iota_{K^2} \end{pmatrix}}_{S^\star} \begin{pmatrix} E \end{pmatrix} \begin{pmatrix} \text{vech}\,(\Phi) \\ \sigma_0^2 \\ \vdots \\ \sigma_{K^2}^2 \end{pmatrix}, \tag{A3}$$

where D_n^+ is the Moore-Penrose inverse of D_n, which is equal to $D_n^+ = (D_n' D_n)^{-1} D_n'$, see Magnus and Neudecker (2007, p. 57).

The post-multiplication by D_P is used to remove the redundant elements of Φ. Since D_P is a zero-one matrix that contains a single nonzero element in each row, $(\Gamma \otimes \Gamma) D_P$ is a $\left(K^2 \times \frac{P(P+1)}{2}\right)$-matrix, of which each column is either a column of $\Gamma \otimes \Gamma$, or the sum of two columns of this matrix, each column of $\Gamma \otimes \Gamma$ being used only once. Therefore, the full rank of $\Gamma \otimes \Gamma$ implies the full rank of $(\Gamma \otimes \Gamma) D_P$.

The pre-multiplication by D_K^+ removes all the redundant equations due to the symmetry of Ω. Therefore, this operation does not reduce the rank of the overall matrix, implying that $A^\star = \begin{pmatrix} S^\star & E \end{pmatrix}$ is also full rank. □

17. $\text{rank}\,(\Gamma \otimes \Gamma) = \text{rank}\,(\Gamma)^2 = P^2$.

B Jacobians of the variable transformations: Proofs

B1 Identifying restriction R1a

In the expanded model, $\widetilde{\Sigma} = \widetilde{\sigma}_0^2 \iota_K \iota_K' + \text{diag}(\widetilde{\sigma}_1^2, \cdots, \widetilde{\sigma}_K^2)$ is not restricted, while in the restricted model, Σ is constrained such that $\sigma_1^2 = c$, where $c \in \mathbb{R}^+$. The two versions of the model are related through the following transformations of variables:

$$\alpha^2 = \widetilde{\sigma}_1^2/c, \qquad \widetilde{\sigma}_k^2 = \alpha^2 \sigma_k^2, \qquad k = 0, \cdots, K.$$

The Jacobian of the transformation $\widetilde{\Sigma} \to (\Sigma, \alpha^2)$ is the determinant of the matrix of first derivatives of the function that transforms (Σ, α^2) into $\widetilde{\Sigma}$. This matrix is equal to:

$$\begin{array}{c} \\ \widetilde{\sigma}_1^2 \\ \widetilde{\sigma}_0^2 \\ \widetilde{\sigma}_2^2 \\ \vdots \\ \widetilde{\sigma}_K^2 \end{array} \begin{array}{c} \begin{array}{ccccc} \alpha^2 & \sigma_0^2 & \sigma_2^2 & \cdots & \sigma_K^2 \end{array} \\ \begin{pmatrix} c & 0 & 0 & \cdots & 0 \\ \sigma_0^2 & \alpha^2 & 0 & \cdots & 0 \\ \sigma_2^2 & 0 & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ \sigma_K^2 & 0 & \cdots & 0 & \alpha^2 \end{pmatrix} \end{array} \equiv A,$$

and its determinant is equal to $c(\alpha^2)^K$. Thus, the Jacobian of the transformation $\mathcal{J}_{\widetilde{\Sigma} \to (\Sigma, \alpha^2)}$ is proportional to $(\alpha^2)^K$.

B2 Identifying restriction R1b

The restriction implies that $\sigma_1^2 = c - \sum_{k=0, k\neq 1}^{K} \sigma_k^2$, and the matrix of first derivatives corresponding to the transformation can be expressed as:

$$\begin{array}{c} \\ \widetilde{\sigma}_1^2 \\ \widetilde{\sigma}_0^2 \\ \widetilde{\sigma}_2^2 \\ \vdots \\ \widetilde{\sigma}_K^2 \end{array} \begin{array}{c} \begin{array}{ccccc} \alpha^2 & \sigma_0^2 & \sigma_2^2 & \cdots & \sigma_K^2 \end{array} \\ \begin{pmatrix} \sigma_1^2 & -\alpha^2 & -\alpha^2 & \cdots & -\alpha^2 \\ \sigma_0^2 & \alpha^2 & 0 & \cdots & 0 \\ \sigma_2^2 & 0 & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ \sigma_K^2 & 0 & \cdots & 0 & \alpha^2 \end{pmatrix} \end{array} \equiv A,$$

which can be rewritten as:

$$A = \alpha^2 \left(I_{K+1} + \frac{U'V}{\alpha^2} \right),$$

with:

$$U = \begin{pmatrix} \sigma_1^2 - \alpha^2 & \sigma_0^2 & \sigma_2^2 & \cdots & \sigma_K^2 \\ -\alpha^2 & 0 & \cdots & \cdots & 0 \end{pmatrix},$$
$$V = \begin{pmatrix} 1 & 0 & \cdots & \cdots & 0 \\ 0 & 1 & \cdots & \cdots & 1 \end{pmatrix}.$$

The determinant of A can then be computed using some basic properties of the determinant and Sylvester's determinant theorem:

$$\begin{aligned} |A| = (\alpha^2)^{K+1} \left| I_{K+1} + \frac{U'V}{\alpha^2} \right| &= (\alpha^2)^{K+1} \left| I_2 + \frac{VU'}{\alpha^2} \right|, \\ &= (\alpha^2)^{K+1} \left| I_2 + \frac{1}{\alpha^2} \begin{pmatrix} \sigma_1^2 - \alpha^2 & -\alpha^2 \\ \sum_{k=0,k\neq 1}^K \sigma_k^2 & 0 \end{pmatrix} \right|, \\ &= (\alpha^2)^{K+1} \left(\frac{\sum_{k=0}^K \sigma_k^2}{\alpha^2} \right), \\ &= c\,(\alpha^2)^K. \end{aligned}$$

The Jacobian of the transformation is therefore the same as for scheme R1a and is proportional to $(\alpha^2)^K$.

B3 Identifying restriction R2a

In the expanded model, $\widetilde{\Phi}$ and $\widetilde{\Sigma} = \widetilde{\sigma}_0^2 \iota_K \iota_K' + \text{diag}(\widetilde{\sigma}_1^2, \cdots, \widetilde{\sigma}_K^2)$ are not restricted, while in the restricted model, Φ and Σ are constrained such that $\Phi_{[11]} + \sigma_0^2 + \sigma_1^2 = c$. The two versions of the model are related through the following transformations of variables:

$$\alpha^2 = \left(\widetilde{\Phi}_{[11]} + \widetilde{\sigma}_0^2 + \widetilde{\sigma}_1^2 \right) / c, \qquad \widetilde{\Phi}_{[ij]} = \alpha^2 \Phi_{[ij]}, \qquad i,j = 1, \cdots, P,$$
$$\widetilde{\sigma}_k^2 = \alpha^2 \sigma_k^2, \qquad k = 0, \cdots, K.$$

Using the identity $\Phi_{[11]} = c - \sigma_0^2 - \sigma_1^2$ derived from the restriction, the matrix of first

derivatives of the function that transforms (Φ, Σ, α^2) into $(\widetilde{\Phi}, \widetilde{\Sigma})$ is equal to:[18]

$$
\begin{array}{c}
 \\
\widetilde{\Phi}_{[11]} \\
\widetilde{\Phi}_{[12]} \\
\widetilde{\Phi}_{[13]} \\
\vdots \\
\widetilde{\Phi}_{[PP]} \\
\widetilde{\sigma}_0^2 \\
\widetilde{\sigma}_1^2 \\
\vdots \\
\vdots \\
\widetilde{\sigma}_K^2
\end{array}
\begin{array}{c}
\begin{array}{ccccccccccc}
\alpha^2 & \Phi_{[12]} & \Phi_{[13]} & \cdots & \Phi_{[PP]} & \sigma_0^2 & \sigma_1^2 & \sigma_2^2 & \cdots & \sigma_K^2
\end{array} \\
\left(\begin{array}{ccccccccccc}
\Phi_{[11]} & 0 & 0 & \cdots & 0 & -\alpha^2 & -\alpha^2 & 0 & \cdots & 0 \\
\Phi_{[12]} & \alpha^2 & 0 & \cdots & & & & & \cdots & 0 \\
\Phi_{[13]} & 0 & \alpha^2 & \ddots & & & & & & \vdots \\
\vdots & \vdots & \ddots & \ddots & & & & & & \\
\Phi_{[PP]} & & & & \ddots & & & & & \\
\sigma_0^2 & & & & & \ddots & & & & \\
\sigma_1^2 & & & & & & \ddots & & & \\
\vdots & & & & & & & \ddots & \ddots & \vdots \\
\vdots & \vdots & & & & & & \ddots & \ddots & 0 \\
\sigma_K^2 & 0 & \cdots & & & & & \cdots & 0 & \alpha^2
\end{array}\right)
\end{array}
\equiv A.
$$

The matrix A can be expressed as:

$$
A = \alpha^2 \left(I_Q + \frac{U'V}{\alpha^2} \right), \qquad\qquad Q = \frac{P(P+1)}{2} + K + 1,
$$

with:

$$
U = \begin{pmatrix}
\Phi_{[11]} - \alpha^2 & \Phi_{[12]} & \cdots & \Phi_{[PP]} & \sigma_0^2 & \sigma_1^2 & \cdots & \sigma_K^2 \\
-\alpha^2 & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & 0
\end{pmatrix},
$$

$$
V = \begin{pmatrix}
1 & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & 0 \\
0 & 0 & \cdots & 0 & 1 & 1 & 0 & \cdots & 0
\end{pmatrix}.
$$

Similarly to Appendix B2, we obtain:

$$
\begin{aligned}
|A| = (\alpha^2)^Q \left| I_Q + \frac{U'V}{\alpha^2} \right| &= (\alpha^2)^Q \left| I_2 + \frac{VU'}{\alpha^2} \right|, \\
&= (\alpha^2)^Q \left| I_2 + \frac{1}{\alpha^2} \begin{pmatrix} \Phi_{[11]} - \alpha^2 & -\alpha^2 \\ \sigma_0^2 + \sigma_1^2 & 0 \end{pmatrix} \right|, \\
&= (\alpha^2)^Q \left(\frac{\Phi_{[11]} + \sigma_0^2 + \sigma_1^2}{\alpha^2} \right), \\
&= c\,(\alpha^2)^{Q-1}.
\end{aligned}
$$

Therefore, the Jacobian of the transformation $\mathcal{J}_{(\widetilde{\Phi},\widetilde{\Sigma})\to(\Phi,\Sigma,\alpha^2)}$ is proportional to $\left(\alpha^2\right)^{\frac{P(P+1)}{2}+K}$.

18. Alternatively, we could have omitted any of the other two restricted parameters, σ_0^2 or σ_1^2, without changing the result on the Jacobian of the transformation.

B4 Identifying restriction R2b

The proof is similar to the one used for scheme R2a. The restriction $\mathrm{tr}(\Phi) + \sum_{k=0}^{K} \sigma_k^2 = c$ implies that $\Phi_{[11]} = c - \sum_{j=2}^{P} \Phi_{[jj]} - \sum_{k=0}^{K} \sigma_k^2$. The matrix of first derivatives is equal to:

$$
\begin{array}{c}
 \\
\widetilde{\Phi}_{[11]} \\
\widetilde{\Phi}_{[22]} \\
\widetilde{\Phi}_{[33]} \\
\vdots \\
\widetilde{\Phi}_{[PP]} \\
\widetilde{\Phi}_{[12]} \\
\vdots \\
\widetilde{\Phi}_{[1P]} \\
\widetilde{\sigma}_0^2 \\
\vdots \\
\vdots \\
\widetilde{\sigma}_K^2
\end{array}
\begin{array}{c}
\begin{array}{cccccccccccc}
\alpha^2 & \Phi_{[22]} & \Phi_{[33]} & \cdots & \Phi_{[PP]} & \Phi_{[12]} & \cdots & \Phi_{[1P]} & \sigma_0^2 & \sigma_1^2 & \cdots & \sigma_K^2
\end{array} \\
\left(\begin{array}{cccccccccccc}
\Phi_{[11]} & -\alpha^2 & -\alpha^2 & \cdots & -\alpha^2 & 0 & \cdots & 0 & -\alpha^2 & -\alpha^2 & \cdots & -\alpha^2 \\
\Phi_{[12]} & \alpha^2 & 0 & \cdots & & & & & & & \cdots & 0 \\
\Phi_{[13]} & 0 & \alpha^2 & \ddots & & & & & & & & \vdots \\
\vdots & \vdots & \ddots & \ddots & & & & & & & & \\
\Phi_{[PP]} & & & & \ddots & & & & & & & \\
\Phi_{[12]} & & & & & \ddots & & & & & & \\
\vdots & & & & & & \ddots & & & & & \\
\Phi_{[1P]} & & & & & & & \ddots & & & & \\
\sigma_0^2 & & & & & & & & \ddots & & & \\
\vdots & & & & & & & & & \ddots & \ddots & \vdots \\
\vdots & \vdots & & & & & & & & \ddots & \ddots & 0 \\
\sigma_K^2 & 0 & \cdots & & & & & & & \cdots & 0 & \alpha^2
\end{array}\right)
\end{array}
\equiv A,
$$

which can be expressed as:

$$
A = \alpha^2 \left(I_Q + \frac{U'V}{\alpha^2} \right), \qquad\qquad Q = \frac{P(P+1)}{2} + K + 1,
$$

with:

$$
U = \begin{pmatrix}
\Phi_{[11]} - \alpha^2 & \Phi_{[12]} & \cdots & \Phi_{[PP]} & \sigma_0^2 & \sigma_1^2 & \cdots & \sigma_K^2 \\
-\alpha^2 & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & 0
\end{pmatrix},
$$

$$
V = \begin{pmatrix}
1 & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & 0 \\
0 & 1 & \cdots & 1 & 0 & \cdots & 0 & 1 & \cdots & 1
\end{pmatrix}.
$$

Using again Sylvester's determinant theorem, it comes:

$$
\begin{aligned}
|A| = (\alpha^2)^Q \left| I_Q + \frac{U'V}{\alpha^2} \right| &= (\alpha^2)^Q \left| I_2 + \frac{VU'}{\alpha^2} \right|, \\
&= (\alpha^2)^Q \left| I_2 + \frac{1}{\alpha^2} \begin{pmatrix} \Phi_{[11]} - \alpha^2 & -\alpha^2 \\ \sum_{j=2}^{P} \Phi_{[jj]} + \sum_{k=0}^{K} \sigma_k^2 & 0 \end{pmatrix} \right|, \\
&= (\alpha^2)^Q \left(\frac{\sum_{j=1}^{P} \Phi_{[jj]} + \sum_{k=0}^{K} \sigma_k^2}{\alpha^2} \right), \\
&= c\,(\alpha^2)^{Q-1},
\end{aligned}
$$

thus showing that the Jacobian of the transformation is the same as in scheme R2a and is proportional to $\left(\alpha^2\right)^{\frac{P(P+1)}{2}+K}$.

C Details of the MCMC Sampler

C1 Step 1: Sampling the latent utilities jointly with the working parameter

Since $p(\widetilde{Y}, \alpha^2 \mid \beta, \Phi, \Sigma) \propto p(\widetilde{Y} \mid \alpha^2, \beta, \Phi, \Sigma) p(\alpha^2 \mid \Phi, \Sigma)$, the sampling is done by first sampling α^2 from its prior specified in Eq. (21), then drawing Y from its conditional distribution given β, Φ and Σ, and finally transforming $\widetilde{Y} = \alpha Y$.

The latent utilities are sampled from a multivariate truncated normal distribution with covariance matrix $\Gamma\Phi\Gamma' + \Sigma$ using a standard sampling scheme, as for instance in McCulloch and Rossi (1994, Section 3).

C2 Step 2: Sampling the regression coefficients jointly with the working parameter

Sampling from $p(\widetilde{\beta}, \alpha^2 \mid \widetilde{Y}, \Phi, \Sigma)$ is done in two steps by first sampling the working parameter from its marginalized conditional distribution $p(\alpha^2 \mid \widetilde{Y}, \Phi, \Sigma)$, then conditional on this draw by sampling the regression coefficients from $p(\widetilde{\beta} \mid \alpha^2, \widetilde{Y}, \Phi, \Sigma)$.

Given the prior distribution on the regression coefficients implied in the expanded model (see Eq. (29)), the conditional distribution of $\widetilde{\beta}$ is:

$$
\widetilde{\beta} \mid \alpha^2, \widetilde{Y}, \Phi, \Sigma \sim \mathcal{N}\left(B_\beta b_\beta;\, \alpha^2 B_\beta\right), \qquad B_\beta^{-1} = B_0^{-1} + \sum_{i=1}^{N} X_i'(\Gamma\Phi\Gamma' + \Sigma)^{-1} X_i,
$$

$$
b_\beta = \sum_{i=1}^{N} X_i'(\Gamma\Phi\Gamma' + \Sigma)^{-1} \widetilde{Y}_i.
$$

The conditional distribution of the working parameter is derived as a function of the likelihood function, the prior and posterior distributions of $\widetilde{\beta}$, and the prior of α^2. It can be evaluated at any value of $\widetilde{\beta}$, for instance at the posterior mean of $\widetilde{\beta}$ to simplify

calculations, i.e., at $\widehat{\beta} = B_\beta b_\beta$:

$$p(\alpha^2 \mid \widetilde{Y}, \Phi, \Sigma) \propto \left. \frac{p(\widetilde{Y} \mid \alpha^2, \widetilde{\beta}, \Phi, \Sigma) p(\widetilde{\beta} \mid \alpha^2) p(\alpha^2 \mid \Phi, \Sigma)}{p(\widetilde{\beta} \mid \widetilde{Y}, \alpha^2, \Phi, \Sigma)} \right|_{\widetilde{\beta}=\widehat{\beta}}, \tag{C1}$$

where:

$$\begin{aligned}
p(\widetilde{Y} \mid \alpha^2, \widetilde{\beta}, \Phi, \Sigma) &\propto \left|\alpha^2(\Gamma\Phi\Gamma' + \Sigma)\right|^{-N/2} \\
&\quad \times \exp\left\{ -\frac{1}{2\alpha^2} \sum_{i=1}^{N} (\widetilde{Y}_i - X_i\widetilde{\beta})'(\Gamma\Phi\Gamma' + \Sigma)^{-1}(\widetilde{Y}_i - X_i\widetilde{\beta}) \right\}, \\
p(\widetilde{\beta} \mid \alpha^2) &\propto \left|\alpha^2 B_0\right|^{-1/2} \exp\left\{ -\frac{1}{2\alpha^2} \widetilde{\beta}' B_0^{-1} \widetilde{\beta} \right\}, \\
p(\widetilde{\beta} \mid \widetilde{Y}, \alpha^2, \Phi, \Sigma) &\propto \left|\alpha^2 B_\beta\right|^{-1/2} \exp\left\{ -\frac{1}{2\alpha^2} (\widetilde{\beta} - B_\beta b_\beta)' B_\beta^{-1} (\widetilde{\beta} - B_\beta b_\beta) \right\}.
\end{aligned}$$

The working parameter α^2 has two different conditional prior distributions in the four identifying restriction schemes, which provides two different posterior distributions.

For the schemes R1a and R1b, since:

$$p(\alpha^2 \mid \Phi, \Sigma) = p(\alpha^2 \mid \Sigma) \propto \left(\alpha^2\right)^{-a_0(K+1)-1} \exp\left\{ -\frac{b_0}{\alpha^2} \sum_{k=0}^{K} \frac{1}{\sigma_k^2} \right\},$$

Eq. (C1) evaluated at $\widehat{\beta}$ provides:

$$p(\alpha^2 \mid \widetilde{Y}, \Phi, \Sigma) \propto \left(\alpha^2\right)^{-a_0(K+1)-KN/2-1} \exp\left\{ -\frac{\tau_\alpha}{\alpha^2} \right\},$$

with:

$$\tau_\alpha = \frac{1}{2} \sum_{i=1}^{N} (\widetilde{Y}_i - X_i\widehat{\beta})'(\Gamma\Phi\Gamma' + \Sigma)^{-1}(\widetilde{Y}_i - X_i\widehat{\beta}) + \frac{1}{2}\widehat{\beta}' B_0^{-1} \widehat{\beta} + b_0 \sum_{k=0}^{K} \frac{1}{\sigma_k^2},$$

which corresponds to the kernel of an inverse-Gamma distribution:

$$\alpha^2 \mid \widetilde{Y}, \Phi, \Sigma \sim \mathcal{G}^{-1}(a_0(K+1) + KN/2;\, \tau_\alpha).$$

For the schemes R2a and R2b, since

$$p(\alpha^2 \mid \Phi, \Sigma) \propto \left(\alpha^2\right)^{-(\nu_0 P + 2a_0(K+1))/2-1} \exp\left\{ -\frac{t_0}{2\alpha^2} \left[\operatorname{tr}\left(S_0(\Phi)^{-1}\right) + 2b_0 \sum_{k=0}^{K} \frac{1}{\sigma_k^2} \right] \right\},$$

Eq. (C1) evaluated at $\widehat{\beta}$ provides:

$$p(\alpha^2 \mid \widetilde{Y}, \Phi, \Sigma) \propto \left(\alpha^2\right)^{-\delta_\alpha/2-1} \exp\left\{ -\frac{\tau_\alpha}{2\alpha^2} \right\},$$

with:

$$\delta_\alpha = NK + \nu_0 P + 2a_0(K+1),$$
$$\tau_\alpha = \sum_{i=1}^{N} (\widetilde{Y}_i - X_i\widehat{\beta})'(\Gamma\Phi\Gamma' + \Sigma)^{-1}(\widetilde{Y}_i - X_i\widehat{\beta}) + \widehat{\beta}' B_0^{-1}\widehat{\beta} + t_0 \mathrm{tr}\big(S_0(\Phi)^{-1}\big) + 2t_0 b_0 \sum_{k=0}^{K} \frac{1}{\sigma_k^2},$$

which corresponds to the kernel of a scaled inverse chi-squared distribution:

$$\alpha^2 \mid \widetilde{Y}, \Phi, \Sigma \sim \tau_\alpha / \chi^2_{(\delta_\alpha)}.$$

C3 Step 3: Sampling the specific effects and the baseline error term

These latent variables are sampled in the expanded model by first drawing $\widetilde{\eta}$ from $p(\widetilde{\eta} \mid \widetilde{Y}, \widetilde{\beta}, \Phi, \Sigma, \alpha^2)$, then by drawing $\widetilde{u}_0$ from $p(\widetilde{u}_0 \mid \widetilde{\eta}, \widetilde{Y}, \widetilde{\beta}, \Phi, \Sigma, \alpha^2)$, using the value of α^2 sampled in step 2.

The differenced system in Eq. (8) can be seen, for each individual $i = 1, \ldots, N$ of the sample, as a linear regression model where the specific effects η_i represent the regression coefficients associated with the matrix of indicators Γ. Therefore, in the expanded model the specific effects are sampled from:

$$\widetilde{\eta}_i \mid \widetilde{Y}_i, \widetilde{\beta}, \Phi, \Sigma, \alpha^2 \sim \mathcal{N}\big(B_\eta b_{\eta_i};\, \alpha^2 B_\eta\big), \qquad B_\eta^{-1} = \Gamma'(\Sigma)^{-1}\Gamma + (\Phi)^{-1}, \quad b_{\eta_i} = \Gamma'(\Sigma)^{-1}(\widetilde{Y}_i - X_i\widetilde{\beta}).$$

Similarly, the baseline error term $\widetilde{u}_0$ can be seen as the regression coefficient of a column vector of K minus ones in Eq. (8). Consequently, it is sampled as:[19]

$$\widetilde{u}_{0i} \mid \widetilde{\eta}_i, \widetilde{Y}_i, \widetilde{\beta}, \Sigma, \alpha^2 \sim \mathcal{N}\big(B_{u_0} b_{u_{0i}};\, \alpha^2 B_{u_0}\big), \quad B_{u_0}^{-1} = \sum_{k=0}^{K} \frac{1}{\sigma_k^2}, \quad b_{u_{0i}} = -\left(\frac{1}{\sigma_1^2}, \ldots, \frac{1}{\sigma_K^2}\right)(\widetilde{Y}_i - X_i\widetilde{\beta} - \Gamma\widetilde{\eta}_i),$$

where the minus sign at the beginning of $b_{u_{0i}}$ comes from the fact that the baseline error term $\widetilde{u}_0$ is substracted from the remaining error terms.

C4 Step 4: Sampling the covariance matrix of the specific effects and the idiosyncratic variances jointly with the working parameter

This step is slightly different for R1* and R2*, as the prior distribution on $\widetilde{\Phi}$ depends on the working parameter α^2 in R1*, but not in R2*. Since $p(\widetilde{\Phi}, \widetilde{\Sigma}, \alpha^2 \mid \widetilde{Y}, \widetilde{\beta}, \widetilde{\eta}, \widetilde{u}_0) = p(\widetilde{\Phi} \mid \alpha^2, \widetilde{\eta}) p(\widetilde{\Sigma}, \alpha^2 \mid \widetilde{Y}, \widetilde{\beta}, \widetilde{\eta}, \widetilde{u}_0)$, both schemes start by sampling the idiosyncratic variances,

19. Note that Φ is dropped from the conditioning set, as $\widetilde{\eta}$ is already conditioned upon.

expressed as $\widetilde{\Sigma} = \widetilde{\sigma}_0^2 \iota_K \iota_K' + \text{diag}(\widetilde{\sigma}_1^2, \ldots, \widetilde{\sigma}_K^2)$, in the expanded model:

$$\widetilde{\sigma}_0^2 \mid \widetilde{u}_0 \sim \mathcal{G}^{-1}\left(a_0 + \frac{N}{2};\, t_0 b_0 + \frac{1}{2}\sum_{i=1}^{N} (\widetilde{u}_{0i})^2\right),$$
$$\widetilde{\sigma}_k^2 \mid \widetilde{Y}, \widetilde{\beta}, \widetilde{\eta}, \widetilde{u}_0 \sim \mathcal{G}^{-1}\left(a_0 + \frac{N}{2};\, t_0 b_0 + \frac{1}{2}\sum_{i=1}^{N} (\widetilde{Y}_{ki} - X_{ki}\widetilde{\beta} - \Gamma_k \widetilde{\eta}_i + \widetilde{u}_{0i})^2\right),$$

for $k = 1, \ldots, K$, where $\widetilde{Y}_{ki}$, X_{ki}, Γ_k denote the kth element/row of the corresponding vector/matrices $\widetilde{Y}_i$, X_i, Γ, respectively, and where $t_0 = 1$ for the two R1* schemes.

For the first two sampling schemes, the working parameter is retrieved as $\alpha^2 = \widetilde{\sigma}_1^2 / c$ (scheme R1a) or as $\alpha^2 = \left[\sum_{k=0}^{K} \widetilde{\sigma}_k^2\right] / c$ (scheme R1b). Conditional on the value of the working parameter, the covariance matrix of the specific effects $\widetilde{\Phi}$, with prior distribution specified in Eq. (30), is sampled from:

$$\widetilde{\Phi} \mid \alpha^2, \widetilde{\eta} \sim \mathcal{W}^{-1}\big(\nu_0 + N;\, \widetilde{\eta}'\widetilde{\eta} + \alpha^2 S\big).$$

In sampling schemes R2a and R2b, the prior distribution of $\widetilde{\Phi}$ does not depend on the working parameter in the expanded model. Therefore, the inverse-Wishart prior distribution assumed in Eq. (24) implies that:

$$\widetilde{\Phi} \mid \widetilde{\eta} \sim \mathcal{W}^{-1}\big(\nu_0 + N;\, \widetilde{\eta}'\widetilde{\eta} + t_0 S\big).$$

Given the sampled values of $\widetilde{\Sigma}$ and $\widetilde{\Phi}$, the working parameter is then retrieved as $\alpha^2 = \left[\widetilde{\Phi}^{\star}_{[1,1]} + \widetilde{\sigma}_0^2 + \widetilde{\sigma}_1^2\right] / c$ (scheme R2a) or as $\alpha^2 = \left[\text{tr}\left(\widetilde{\Phi}\right) + \sum_{k=0}^{K} \widetilde{\sigma}_k^2\right] / c$ (scheme R2b).