A Multinomial Probit Model with Latent Factors

Identification and Interpretation without a Measurement System

Rémi Piatek*

Department of Economics
University of Copenhagen

Miriam Gensowski

Department of Economics
University of Copenhagen
and IZA

Abstract

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 a structure on a few drivers of unobserved heterogeneity *a priori*. One such example is the choice among 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.

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

*Corresponding author: Department of Economics, University of Copenhagen, Øster Farimagsgade 5, 1353 Copenhagen K, Denmark. Tel: (+45) 35 32 30 35. E-mail: remi.piatek@econ.ku.dk.

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.

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), random intercepts and random coefficients (Haaijer et al., 1998), 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 improving 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 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 model 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 and Meng, 2001). We implement these techniques and construct suitable working parameters that build on our identification restrictions.

We apply our framework to both synthetic and real data. In an extensive Monte Carlo study, we compare our methodology to the benchmark MNP with a plain covariance matrix, on both a modestly sized model with six alternatives (like our motivating example), and a larger one with 16 alternatives. We demonstrate that our approach performs at least as well as the state-of-the-art MNP model, and outperforms it in recovering the covariance structure of the model in larger settings. The simulations also show that our sampler manages to divide the unobservables of the model into interpretable latent factors and idiosyncratic noise. In this respect, our approach provides an advantage in terms of interpretability and economic content of the model. That being said, the simulations also highlight that some variants of our approach, as defined by the specific identification restrictions they rely on, allow stronger empirical identification than others. We conclude with a recommendation to practitioners on which identification restriction to prefer in practice.

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. This example is estimated on real data from the National Longitudinal Survey of Youth '79 (Bureau of Labor Statistics, U.S. Department of Labor, 2014), where we illustrate how our approach can serve to study the latent taste shocks, their correlation, and marginal effects.

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 $i = 1, \dots, N$ chooses one of $K+1$ alternatives $D_i \in \{0, 1, \dots, K\}$ 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 $k = 0, 1, \dots, K$, 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

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).

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 $\Gamma^{\star}$:

$$\varepsilon_i^{\star} = \Gamma^{\star}\, \eta_i^{\star} + u_i, \tag{3}$$

where $\varepsilon_i^{\star} = (\varepsilon_{i0}^{\star}, \ldots, \varepsilon_{iK}^{\star})'$, $\Gamma^{\star}$ is a user-specified matrix of dimension $(K+1) \times J$, $\eta_i^{\star} = (\eta_{i1}^{\star}, \ldots, \eta_{iJ}^{\star})'$ is the J-vector of latent factors, and $u_i = (u_{i0}, \ldots, u_{iK})'$.

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 $\Sigma^{\star} = \operatorname{diag}(\sigma_0^2, \ldots, \sigma_K^2)$ 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:

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

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 a 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 $\eta_{ij}^{\star}$ into the latent utilities. We impose this fixed structure for two reasons. The first one 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 $\eta_i^\star$ 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 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

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.

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 types of 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, $N_O = 3$, 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 their possible combinations. The taste shocks (factors) are allowed to correlate, with the following partitioned covariance matrix:

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

3. In our empirical application in Section 5, "occupation" is not meant as a precise job, but rather as 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.

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

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 $\Delta_K = \begin{pmatrix} -\iota_K & I_K \end{pmatrix}$, where ι_K denotes the vector of ones of length K, and I_K is the identity matrix of dimension K:

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

with:

$$X_i = \Delta_K W_i, \qquad \omega_i = \Delta_K u_i. \tag{9}$$

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

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.

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.

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 $P = \operatorname{rank}(\Delta_K \Gamma^\star)$, 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:

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

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

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

$$\Gamma = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix}. \tag{15}$$

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 two covariance matrices Ω and $\check{\Omega}$ fulfilling Eq. (14) for both pairs exist, 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., $\text{rank}\,(\Gamma) = P$.

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. □

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.

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:

$$\underset{(K\times P)}{\overline{\Gamma}} = \begin{pmatrix} I_P \\ \iota_{(N_S-1)} \otimes I_{(N_O-1)} \quad I_{(N_S-1)} \otimes \iota_{(N_O-1)} \end{pmatrix} \equiv \begin{pmatrix} I_P \\ \overline{\Gamma}_2 \end{pmatrix} \tag{16}$$

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.

Restrictions	Idiosyncratic variances	Overall covariance matrix
Single element	**R1a** $\sigma_1^2 = c$	**R2a** $\Omega_{[1,1]} \equiv \Phi_{[1,1]} + \sigma_0^2 + \sigma_1^2 = c$
Diagonal elements	**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

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.

in a simple way, as it only restricts a single parameter. All the 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. As will be discussed later, 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 plain 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.

3.1 Principle

Marginal Data Augmentation (MDA) methods have been introduced by Meng and van Dyk (1999) and van Dyk and Meng (2001), and developed in parallel with Parameter-Expansion methods (C. Liu et al., 1998; J. S. 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

8. Package available from the authors upon request.

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:

$$\begin{aligned}\mathcal{L}(\beta, \Phi, \Sigma \mid D, X) &\propto p(D \mid X, \beta, \Phi, \Sigma), \\ &\propto \iint p(D, Y, \eta \mid X, \beta, \Phi, \Sigma) dY d\eta, \qquad (17) \\ &\propto \iint \left\{ \int p(D, Y, \eta \mid X, \beta, \Phi, \Sigma, \alpha^2) p(\alpha^2 \mid \Phi, \Sigma) d\alpha^2 \right\} dY d\eta. \qquad (18)\end{aligned}$$

In Eq. (17), 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. (18), 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 $\alpha \in \mathbb{R}^+$:

$$\alpha Y \equiv \widetilde{Y}_i = X_i\widetilde{\beta} + \Gamma\widetilde{\eta}_i + \widetilde{\omega}_i, \qquad \begin{aligned} \widetilde{\eta}_i &\sim \mathcal{N}\left(0;\, \widetilde{\Phi}\right), \\ \widetilde{\omega}_i &\sim \mathcal{N}\left(0;\, \widetilde{\Sigma}\right), \end{aligned} \tag{19}$$

with:

$$\widetilde{\beta} \equiv \alpha\beta, \qquad \widetilde{\eta}_i \equiv \alpha\eta_i, \qquad \widetilde{\omega}_i \equiv \alpha\omega_i, \qquad \widetilde{\Phi} \equiv \alpha^2\Phi, \qquad \widetilde{\Sigma} \equiv \alpha^2\Sigma, \tag{20}$$

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. (20). 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 to average the likelihood over a prior distribution for the working parameter $p(\alpha^2 \mid \Phi, \Sigma)$ in Eq. (18). 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,*

$$\widetilde{\sigma}_k^2 \sim \mathcal{G}^{-1}(a_0;\, b_0)\,, \qquad\qquad a_0, b_0 > 0, \tag{21}$$

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

$$\alpha^2 \mid \sigma_0^2, \ldots, \sigma_K^2 \sim \mathcal{G}^{-1}\left(a_0(K+1);\, b_0 \sum_{k=0}^{K} \frac{1}{\sigma_k^2}\right). \tag{22}$$

Corollary 3.1 (**Marginal prior distribution of Σ in scheme R1a**). *Given the priors in Eqs. (21) and (22), the marginal prior distribution of the idiosyncratic variances in scheme R1a is proportional to:*

$$p(\sigma_0^2, \ldots, \sigma_K^2) \propto \left(\sum_{k=0}^{K} \frac{1}{\sigma_k^2}\right)^{-a_0(K+1)} \prod_{k=0}^{K} (\sigma_k^2)^{-a_0-1}\, \mathbb{1}\{\sigma_1^2 = c\}\,. \tag{23}$$

Proof. See Appendix B1. □

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. (21), 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. (22) and Eq. (23), 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 B3 and B4. □

The kernel in Eq. (23) 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. (21), 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 $\widetilde{\sigma}_0^2, \dots, \widetilde{\sigma}_K^2$. 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 $k = 0, \dots, K$, 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:*

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

Proof. See Appendix B2. □

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 B5 and B6. □

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. (20)) 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:

$$\beta \sim \mathcal{N}(0;\, B_0)\,, \qquad\qquad \widetilde{\beta} \mid \alpha^2 \sim \mathcal{N}\big(0;\, \alpha^2 B_0\big)\,, \tag{28}$$

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:

$$\Phi \sim \mathcal{W}^{-1}(\nu_0;\, S_0)\,, \qquad\qquad \widetilde{\Phi} \mid \alpha^2 \sim \mathcal{W}^{-1}\big(\nu_0;\, \alpha^2 S_0\big)\,. \tag{29}$$

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 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[9] 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

9. See Eq. (13).

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

Sample Y_{ik} from $p(Y_{ik} \mid D_i, Y_{i,-k}, \beta^{(t)}, \Phi^{(t)}, \Sigma^{(t)})$, for $i = 1, \ldots, N$ and $k = 1, \ldots, K$, then set $\widetilde{Y} = \alpha^{(a)} Y$.

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

where $\widetilde{Z} = (\widetilde{Z}_1, \ldots, \widetilde{Z}_N)'$ with $\widetilde{Z}_i = \widetilde{Y}_i - \alpha^{(b)} X_i \beta^{(t+1)} - \Gamma \widetilde{\eta}_i + \widetilde{u}_{0i}$, for $i = 1, \ldots, N$,

using the following accept-reject procedure:

repeat

(a) Sample $\widetilde{\Sigma}$ from $p(\widetilde{\Sigma} \mid \widetilde{Z}, \widetilde{u}_0)$. ▷ *Appendix C4*

(b) Sample $\widetilde{\Phi}$ from $p(\widetilde{\Phi} \mid \widetilde{\eta})$.

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

(d) Compute $Y_i = \left(\widetilde{Z}_i + \alpha^{(c)} X_i \beta^{(t+1)} + \frac{\alpha^{(c)}}{\alpha^{(b)}} \left(\Gamma \widetilde{\eta}_i - \widetilde{u}_{0i}\right)\right) / \alpha^{(c)}$,

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

$$\begin{cases} \max\{Y_{i1}, \ldots, Y_{iK}\} < 0 & \text{if } D_i = 0, \\ \max\{0, Y_{i1}, \ldots, Y_{iK}\} = Y_{ik} & \text{if } D_i = k. \end{cases} \tag{30}$$

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

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.[10] 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 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 $\widetilde{Z}_i = \widetilde{Y}_i - \alpha^{(b)} X_i \beta^{(t+1)} - \Gamma \widetilde{\eta}_i + \widetilde{u}_{0i}$, to take into account the information

10. 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.

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. (30). 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 relying on two different models. The first one has six alternatives and follows the joint education-occupation decision example introduced in Section 2.2, which is the backbone of our empirical illustration presented in Section 5. The second one is a larger model with sixteen alternatives and only four latent factors explaining the unobserved correlation between the latent utilities. In this framework, the factors make it possible to considerably reduce the parameter space—only 10 parameters of interest in the covariance matrix of the factors, compared to 105 off-diagonal elements in the covariance matrix of the error terms in the differenced system of the standard MNP model.

4.1 Experimental setup

Data generation. Synthetic data are generated using the differenced model specified in Eq. (8) combined with the decision rule in Eq. (7). Two models with 6 and 16 alternatives, i.e., $K = 5$ and 15 in the differenced system, respectively, are specified with $P = 3$ and $P = 4$ latent factors factors, respectively. These factors are assumed to be correlated, with covariance matrix Φ specified as a symmetric Toeplitz matrix with elements $(0.5, -0.3, 0.2)$ in the small model, and $(0.5, -0.3, 0.2, 0.1)$ in the larger one.[11] The factors are allocated to the latent utilities through the allocation matrix in Eq. (15) for the small model. In the larger model, the four factors affect the latent utilities in all possible combinations ($2^4 = 16$

11. A symmetric Toeplitz with elements (a, b, c) is defined as $\begin{pmatrix} a & b & c \\ b & a & b \\ c & b & a \end{pmatrix}$.

alternatives) through the following allocation matrix:

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

For the idiosyncratic errors, each variance is fixed to $\sigma_k^2 = 0.25$, for $k = 1, \ldots, K$, implying that $\Sigma = 0.25(I_K + \iota_K \iota_K')$. A single explanatory variable $X_{ik} \overset{iid}{\sim} \mathcal{N}(-0.6; 1)$, which varies across alternatives $k = 1, \ldots, K$ and individuals $i = 1, \ldots, N$, is specified in each model. It influences the latent utilities through the regression coefficient $\beta = 1$ in both cases.

These parametrizations result in samples with a well-balanced number of observations across the different alternatives. We generate data sets of sizes $N = 1,000$, $5,000$ and $10,000$ observations for the two models.

Identification and prior specification. 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 as implemented in the `MNP` package (Imai and van Dyk, 2005b) using both the single-element restriction (Imai and van Dyk, 2005a, we call it MNPa) and the trace restriction (Burgette and Nordheim, 2012, called MNPb).

For the prior distribution, we use $B_0 = 100$ for the variance of the regression coefficient, $a_0 = 2$ and $b_0 = 1$ for the shape and scale of the idiosyncratic variances, and $\nu_0 = P + 2$ for the degrees of freedom of the covariance matrix of the factors.[12] For the scale matrix of the covariance matrix of the factors, we use $S_0 = 2I_P$ in the R1a case, $S_0 = s_0 I_P$ in the R1b case, with $s_0 = 0.2$ in the small model and $s_0 = 0.05$ in the large model, and $S_0 = 3I_P$ in the two R2* cases. The tuning parameter t_0 is set to 1. In each of the four cases, the constant c determining the scale of the latent utilities is fixed using the true values of the parameters to imply $\alpha^2 = 1$, see Table 2.

The prior specification we use is rather non-informative, and similar across the four identification strategies with regard to the weight put on the share of the overall variance of the latent utilities explained by the latent factors. This is achieved through the adjustment of the scale matrix S_0, which is different in the different cases. Alternatively, it would be

12. Setting the number of degrees of freedom to $\nu_0 = P + 1$ would imply a uniform marginal distribution of the correlations between the factors in the expanded model (Barnard et al., 2000). Increasing ν_0, as we do here, pulls the correlations away from ± 1, thus preventing extreme cases with highly-correlated latent factors.

possible to tune the prior of the idiosyncratic variances to come to a similar result. Indeed, 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. As a consequence, it is important to study the role of the prior parameters of the unobserved components of the model together.

A note about side-by-side comparisons of the results from our four identification schemes is in order. As explained in Section 3.3, each of the identification strategies will have a different prior distribution induced even with a single specification of prior parameters, because of the specific parameter transformations used to obtain each identified model. Therefore, our posterior results should be contrasted with caution, 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.

For the benchmark standard MNP model, we specify $B_0 = 100$, $\nu_0 = P + 2$, and $S_0 = I_P$.

Monte Carlo setup and MCMC tuning. For each model and each sample size, we simulate 100 different data sets using different seeds of the random number generator. The Monte Carlo experiment is then carried out on these data sets for each identification strategy of our approach, and each of the two benchmark MNP approaches.[13] In each Monte Carlo replication, 60,000 MCMC iterations are used to update the model, keeping only the last 50,000 ones for posterior inference.

4.2 Simulation results

The simulation results reveal that our sampler not only recovers the parameters well, but provides more information than the benchmark MNP at equal or superior precision. Furthermore, we learn that while all four identification restrictions of our approach yield a theoretically identified model, large disparities between their respective performances prevail. Based on our simulation results, we give a recommendation to practitioners. We finally observe that the decomposition of the covariance matrix works well already on small data sets, as the average posterior is close to the true value.

Recovering the model parameters. The first criterion used to assess the performance of our sampler is the root mean square error (rMSE) of the inferred parameters. Figure 1

13. I.e., we simulate a total number of 600 data sets and run $3,600$ MCMC simulations.

shows boxplots of the distribution of rMSE across the 100 Monte Carlo replication for each approach.[14] For the parameters other than β (which is scalar), the figure shows an aggregate measure corresponding to the median rMSE within each group of parameters Σ, Φ and Ω.[15] As the benchmark MNP model does not allow to decompose the overall covariance matrix, only results for β and Ω are provided in the MNPa and MNPb cases. For a fair comparison, the different approaches should be evaluated 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).

First, we compare the different identification strategies of our approach to each other. A clear pattern of disparities emerges, where strategies based only on the idiosyncratic variances (R1*) perform worse. Especially in the first one that sets $\sigma_1^2 = c$ (R1a), there is considerable variability in the rMSE of the model parameters across Monte Carlo replications. This is the first lesson of this experiment: Although identification strategies relying on a restriction of the idiosyncratic variances provide models that are *theoretically* identified, in practice it is hard to *empirically* identify them. These restrictions appear to be too loose to guarantee the stability of the latent part of the model. On the contrary, approaches imposing the restriction of the overall covariance matrix of the latent part of the model Ω (R2*) provide very satisfactory results.

The second lesson from our simulation study is that approaches R2* always perform at least as well as standard MNP approaches without latent factors. Notably, they outperform the benchmark approach in terms of precision of Ω, especially in larger models—in our example with 16 alternatives, the rMSE of the elements of Ω appears to be up to twice as small in our R2* approaches. Generally, the "trace restriction" (MNPb) and its counterpart in our approach (R1b/R2b) provide better results for all parameters, thus showing that the results of Burgette and Nordheim (2012) for the standard MNP model apply to our approach.

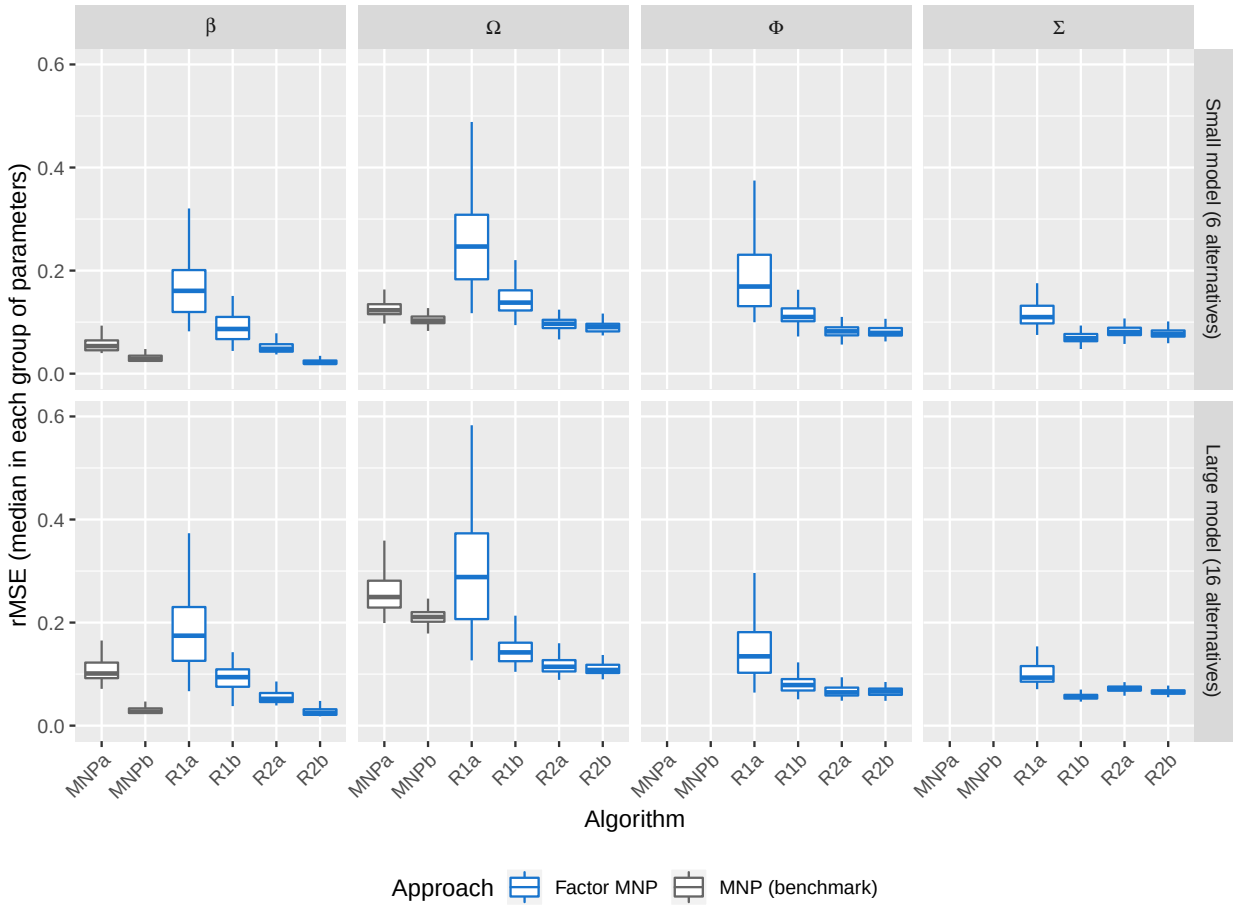
Looking at the decomposition of the unobservables (right two columns of Fig. 1), the same conclusions can be drawn. For Σ and especially Φ, placing the restriction on the idiosyncratic variances Σ is associated with more error than the restriction on the full covariance matrix Ω. The identification restrictions using the full covariance matrix Ω produce much lower rMSEs. Again, the restriction that combines multiple elements of the trace, R2b, works best.

The general lesson of this simulation exercise is that without losing precision, our approach can be used to decompose the otherwise unobserved part of the model, Ω, and that

14. Cases with $N = 1,000$ and $N = 10,000$ observations are reported in Appendix W1. The rMSE is falling in the number of observations, but the comparisons between the variations of our approach and MNP yield the same conclusions.

15. This aggregate measure is only used to limit the number of parameters to be displayed and does not alter the main conclusions of the analysis.

Figure 1: Simulation study — Root mean square error of model parameters, cases with $N = 5,000$ observations.



Notes: Monte Carlo experiment with 100 replications. Box and whiskers plot à la Tukey, where the box shows the 25/50/75th percentiles, and the whiskers 1.5 the inter-quartile range (IQR). Outliers not shown to facilitate the readability of the figure. The statistic of interest is the root mean square error (rMSE) of the parameters, where the median is computed in each group of model parameters for each Monte Carlo replication: Overall covariance matrix of the latent part of the model (Ω), covariance matrix of the latent factors (Φ), idiosyncratic variances of the error terms (Σ). Our approach with the four different restrictions (R1a, R1b, R2a, R2b) is displayed in blue, benchmark MNP model (*without* latent factors, hence no values in panels Φ and Σ) in grey. MNPa: MNP with restriction on first element of the covariance matrix, MNPb: MNP with trace restriction.

this approach can work as well or better than the less-informative MNP if the restriction R2b is used. Through its restriction on a combination of the diagonal elements of Φ and Σ, it is more stable in recovering the parameters and, more importantly, in decomposing the overall variance of the latent utilities into latent factors and error terms. Therefore, a general recommendation is to use R2b in practice.

Separating latent factors from noise. We illustrate the capacity of our sampler 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 $\rho_k = \Gamma_k' \Phi \Gamma_k / (\Gamma_k' \Phi \Gamma_k + \sigma_0^2 + \sigma_k^2)$, for each latent utility $k = 1, \ldots, K$, where Γ_k is the column vector of length P containing the k^{th} row of Γ.

Fig. 2 shows the distribution of the first parameter ρ_1, both *a priori* and *a posteriori*, for each of the two models and across the 100 Monte Carlo replications of the experiment, along with its true value in our synthetic data, for $N = 1,000$, $5,000$ and $10,000$ respectively in the R2b case. 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.

A posteriori, the sampler manages very well to learn from the data the share of overall variance that can be attributed to the latent factors, even on data sets with few observations. The posterior distribution of ρ_1 (blue outline) deviates from the prior distribution in all cases, and the posterior mean (blue triangle) is close to the true value (black dot) already at $N = 1,000$. As the number of observations increase, the posterior distribution simply becomes more concentrated around the true value of the parameter of interest.

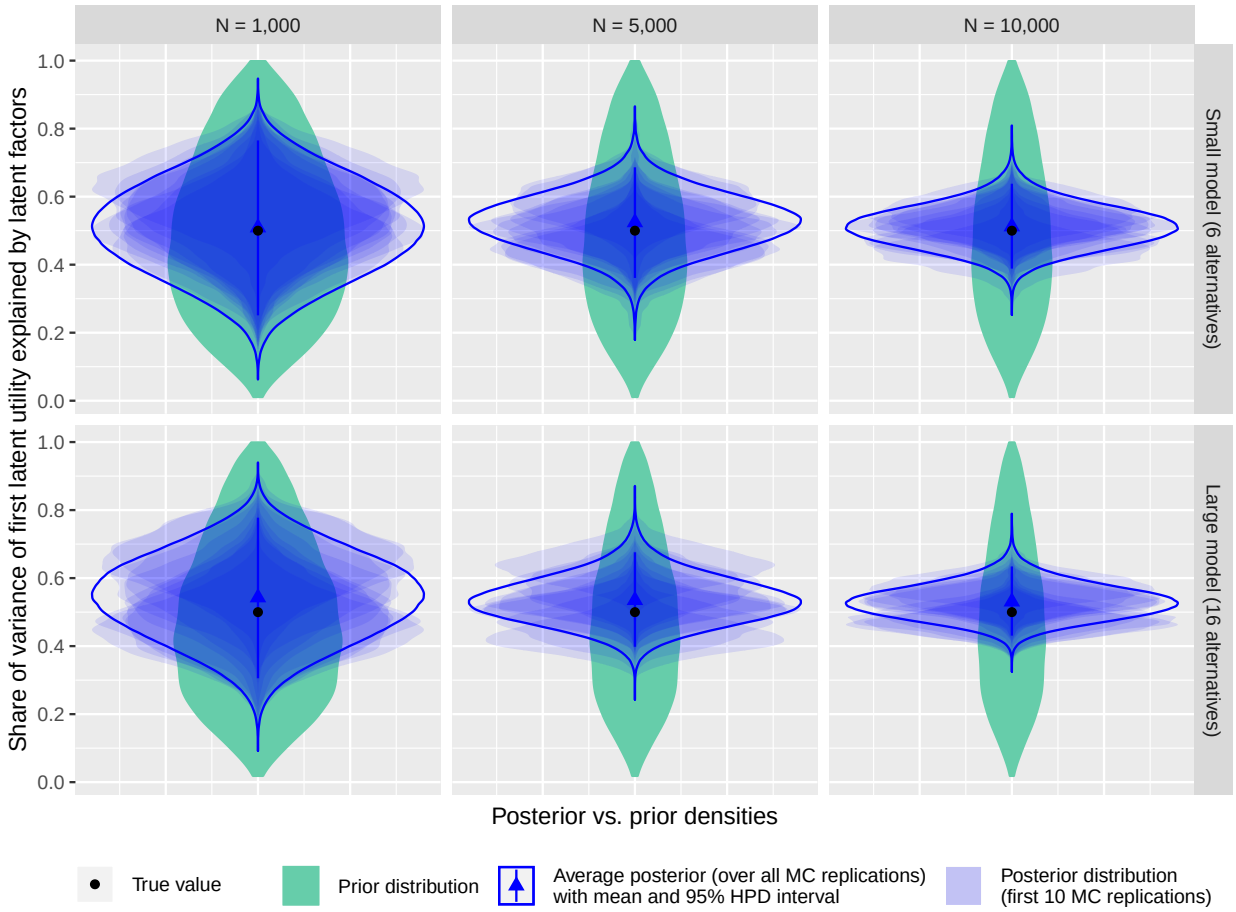
5 Empirical application

We illustrate our approach by confronting the joint educational-occupational decision model introduced in Section 2.2 with real data from the National Longitudinal Survey of Youth '79 (NLSY79, Bureau of Labor Statistics, U.S. Department of Labor, 2014). The sample size and number of alternatives are similar to the small model investigated in our Monte Carlo experiment. In this application, we specifically study the unobserved taste shocks for education and occupation. These latent variables are unraveled by our factor analytical approach, whereas the standard MNP model would not be able to disentangle these effects.

5.1 Data

The NLSY79 followed a sample of American youth born between 1957-64 from 1979 onward and has been used countless times to study education and occupation choices of youth. Thereby, the impact of covariates in our application can be compared to existing studies such as Keane and Wolpin, 1997; Heckman et al., 2006, 2016; Speer, 2017. We use all respondents who have valid information on education and occupation outcomes in the 2010

Figure 2: Simulation study — Proportion of the variance of the first latent utility explained by the latent factors (ρ_1), conditional on the covariates, in the R2b case.



Notes: Average posterior density in dark blue, with mean and 95% highest posterior density (HPD) interval computed over the 100 Monte Carlo replications in each panel. Posterior densities of the first 10 Monte Carlo replications only are displayed in light blue to make the figure better readable. Same prior used for each model (i.e., in each row of the figure). Scales of the violin plots vary across panels.

wave, as well as contemporaneous covariates in 2010 and baseline information from 1979.

The outcome variable is the joint choice of education and occupation, with a total of six pairs. Education can be "Low" with up to 12 years of education (high school graduation), or "High" with some college or more. There are three occupation groups based on 2-digit codes in the census occupational classification system: "Blue collar," "Service" (incl. Sales and Office workers), and "Business/Management/STEM." The resulting six alternatives are listed in Table 3, with baseline level "Low Education/Blue collar."

Following our example in Section 2.2, three latent factors are specified in the differenced version of the model: Two occupation-related factors (η_1 for the taste for "Business" vs. "Blue collar" and η_2 for "Service" vs. "Blue collar"), and a third factor η_3 for the difference

Table 3: Available choices in the NLSY79 survey (2010 and 1979 waves) and allocation of the latent factors to the decisions in our model.

Choice				Factors		
Education	Occupation	Freq	Percent	η_1	η_2	η_3
Low	Blue collar	1,121	18.9	·	·	·
Low	Bus/Management/STEM	435	7.4	1	0	0
Low	Service	1,377	23.3	0	1	0
High	Blue collar	368	6.2	0	0	1
High	Bus/Management/STEM	1,568	26.5	1	0	1
High	Service	1,048	17.7	0	1	1
Total		5,917	100.0			

Note: The baseline in our analysis is "Low education/Blue collar." The latent factors capture differences in tastes for occupations, with baseline "Blue collar" (i.e., η_1 is the difference between the taste for "Business" and "Blue collar", η_2 between "Service" and "Blue collar") and for education, with baseline "Low Education" (i.e., η_3 is the difference in education taste between "High" and "Low").

between the specific taste for high vs. low education. The allocation matrix of the three latent factors, corresponding to Γ in Eq. (15), is given in the rightmost part of Table 3.

For the observed drivers of choices, we select standard covariates that are commonly used in the literature, reflecting the rich possibilities of the NLSY79.[16] We include both alternative-specific covariates (e.g., average local labor market characteristics among workers living in the same region), as well as individual-specific characteristics (e.g., marital status, minority status, willingness to incur risks, all measured in 2010). From the baseline interviews in 1979, we draw on the respondent's performance on the armed forces qualifying test (AFQT, generally used as a proxy for cognitive ability),[17] their mother's highest grade completed, and their occupational expectations (whether they expected to make a career in services or in a blue-collar occupation).

5.2 Inference

We run our version of the MNP model with latent factors using the identification restriction on the overall covariance matrix R2b, as it proved to provide stable results in our Monte Carlo study. For the prior, we specify $B_0 = 10$, $a_0 = 2$, $b_0 = 1$, $\nu_0 = 4$, $S_0 = 4I_4$ and $t_0 = 1$, which is rather non-informative about the latent structure of the model. To check the convergence of our sampler, we run five different Markov chains with different starting

16. Descriptive statistics of the covariates are provided in Table W1.

17. Examples are Neal and Johnson, 1996; Cameron and Heckman, 1998; Altonji and Pierret, 2001; Altonji et al., 2012; Deming, 2017. Note that we use the original AFQT score, not corrected for the endogeneity of schooling on test scores as suggested by Hansen et al., 2004.

values sampled from the prior distribution of the parameters. We use $1,100,000$ MCMC iterations for each chain, and discard the first $100,000$ as burn-in period. The final results are compiled using the MCMC draws from the five chains.

5.3 Empirical results

We first briefly discuss the results that can be drawn from the observed covariates of the model, which should resemble those from a standard MNP model, before turning to the interpretation of the latent part of our factor model approach that allows to go beyond MNP by decomposing the unobservables into latent factors and idiosyncratic errors.

The regression coefficients are all in line with the expectations from the literature (see Table W2 in Appendix W2). Cognitive ability, for example, strongly influences the likelihood of having high educational attainment (as in Cameron and Heckman, 1998), and greater risk-willingness is associated with a lower likelihood of being in a blue-collar occupation (see Dohmen et al., 2011). Workers are more likely to choose an occupation that has higher average wages or a larger share of workers of their own gender.

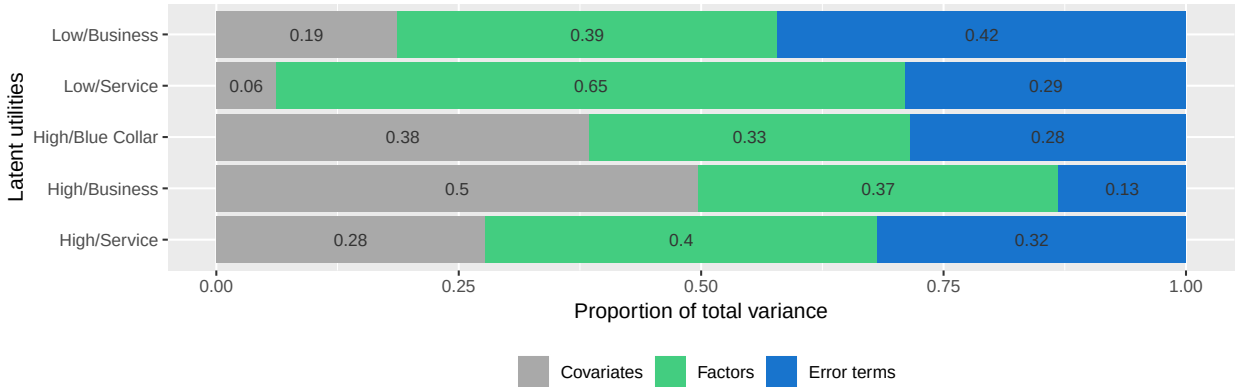
We can now proceed to the added value of our approach relative to the standard MNP. Figure 3 decomposes the variance of the latent utilities to show the respective contributions of the covariates latent factors, and residual error terms. As is often the case in applied work, the covariates do not explain the majority of the variation. For example, only 6% of the variance of the alternative "Low Education/Service" are driven by observable covariates. In a standard MNP framework, the remaining unexplained 94% would simply be lumped together. Our factor structure, however, makes it possible to decompose the unobservables into factors and error terms. In fact, in this particular example the majority of this alternative is driven by specific taste shocks captured by the latent factors.

Referring to the estimation results for the components of Φ in Table 4, we learn that the variance of the latent factor for Service is rather large compared to the other factors. In the case of "Low Education/Service," the corresponding element on the diagonal of Φ is ϕ_{22}.[18] Generally, Fig. 3 shows that the shares of variance explained by latent factors are relatively large, and it demonstrates the importance of occupation or education-specific taste shocks that drive the choice. The data hold information that can be backed out rather than only calling all terms in Ω "noise."

The latent factors not only aid in decomposing the covariance, but they can inform us about choice patterns. Observing the correlations among the factors in Table 4 reveals

18. This decomposition focuses exclusively on the diagonal elements for clarity. The careful reader will notice that the other alternative with service occupations, joint with high education, has a smaller share of variance explained by the latent factor. This is due to the negative covariance between η_2 and η_3.

Figure 3: NLSY79 application — Posterior variance decomposition of the latent utilities.



Notes: Baseline level: Low Education/Blue Collar. Each bar shows the proportion of total variance of the corresponding latent utility explained by the covariates, the latent factors and the error terms, respectively.

more information about the taste shocks. Notably, when comparing two individuals of the same ability, characteristics, and expectations, we see that one who prefers business and management occupations over blue-collar jobs *does not necessarily* also prefer high education. The near-zero result for ϕ_{13} tells us that the taste shocks for business and education are practically uncorrelated. On the other hand, the large and negative ϕ_{23} implies that someone with a preference for service jobs over blue collar tends to dislike high education.

These interpretations of the covariance matrix are informative for the decision-making process, as they go beyond an observation of how frequently choices are made together. The analysis is conditional on both alternative-specific and person-specific covariates, and within this model we can learn about the correlation between taste preferences for the occupation options and education levels separately. 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.

Finally, we can trace out the impact of the latent factors on the choice probabilities. Figure 4 engages in the thought experiment of how changes in the factor η_1, capturing taste shocks for Business vs. Blue collar occupations, affect choice probabilities for a representative individual—all covariates are held constant at their median values in the sample, see Table W1, and the other two factors are integrated out based on the posterior distribution of the model parameters.[19] A stronger taste for business occupations leads to this option being more frequently chosen. But it does not increase this probability evenly: it increases the alternative that combines a Business occupation with *high* education significantly more

19. This kind of figure, which helps understand the marginal effects of the latent factors on the choice probabilities, has been routinely used in the empirical literature (see, e.g., Hansen et al., 2004; Heckman et al., 2006).

Table 4: NLSY79 application — Posterior results for the covariance matrix Φ and corresponding correlation matrix of the latent factors.

	ϕ_{11}	ϕ_{12}	ϕ_{13}	ϕ_{22}	ϕ_{23}	ϕ_{33}
Covariances	0.476	-0.121	0.046	0.828	-0.352	0.469
	(0.178)	(0.187)	(0.131)	(0.177)	(0.126)	(0.169)
Correlations	1.000	-0.198	0.110	1.000	-0.569	1.000
		(0.298)	(0.286)		(0.142)	

Notes: Posterior means and standard deviations (in parentheses) of the covariances and correlations of the latent factors. Factor 1: Taste shock for Business (vs. Blue collar) occupations; Factor 2: Taste shock for Service (vs. Blue collar) occupations; Factor 3: Taste shock for High (vs. Low) education.

so. Furthermore, even though the factor does not directly relate to the relationship between Business and *Service* occupations, the increase in Business stems mostly from decreases in Service occupations. In comparison to the increasingly strong taste for Business jobs, the Service occupations become less attractive—in both education levels. The example highlights that our approach allows to clearly show the complex implications from an increase in a single taste shock that would otherwise remain unobserved. Furthermore, it accentuates once more the practical importance of the structured taste shocks: the choice probability moves from close to zero up to fifty percent, an impressive range.

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. 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 the majority of the "unexplained" variation in utilities.

Figure 4: NLSY79 application — Impact of the latent factor "Business" (vs. "Blue collar") on choice probabilities.

Notes: Probabilities predicted for each ventile of the factor distribution (from 5% to 95%). Covariates fixed at their median values in the sample. 95% highest posterior density intervals in shaded areas. Probabilities simulated based on posterior distribution of model parameters, using a logit-smoothed accept-reject simulator with scale factor 0.1 (see Train, 2009, 5.6).

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 diagonal 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 11$^{\text{th}}$ 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. Special thanks to Sylvia Frühwirth-Schnatter for her constructive suggestions.

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.

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:[20]

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

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 $\text{vec}\,(\Gamma\Phi\Gamma') = (\Gamma\otimes\Gamma)\,\text{vec}\,(\Phi)$, where $\otimes$ denotes the Kronecker product.

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

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

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

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.[21] 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 $\Gamma \otimes \Gamma$.

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,[22] 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. (16)): 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 $M_E(\Gamma \otimes \Gamma)$.

21. The projection matrix M_E is symmetric and idempotent matrix, i.e., $M_E = M_E'$ and $M_E M_E = M_E$.
22. $\text{rank}\,(\Gamma \otimes \Gamma) = \text{rank}\,(\Gamma)^2 = P^2$.

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. □

B Proofs of implied prior distributions and Jacobians

B1 Implied prior distribution in R1a case

This proves Proposition 3.1 and Corollary 3.1:

Proof. The transformation of random variables provides:

$$
\begin{aligned}
p(\alpha^2, \sigma_0^2, \ldots, \sigma_K^2) &= \mathcal{J}_{(\widetilde{\sigma}_0^2,\ldots,\widetilde{\sigma}_K^2)\to(\alpha^2,\sigma_0^2,\ldots,\sigma_K^2)}\, p(\widetilde{\sigma}_0^2, \ldots, \widetilde{\sigma}_K^2), \\
&\propto \left(\alpha^2\right)^K \prod_{k=0}^{K} (\widetilde{\sigma}_k^2)^{-a_0-1} \exp\left\{-\frac{b_0}{\widetilde{\sigma}_k^2}\right\} \mathbb{1}\left\{\sigma_1^2 = c\right\}, \\
&\propto \left(\alpha^2\right)^{-a_0(K+1)-1} \exp\left\{-\frac{b_0}{\alpha^2}\sum_{k=0}^{K}\frac{1}{\sigma_k^2}\right\} \prod_{k=0}^{K} (\sigma_k^2)^{-a_0-1} \mathbb{1}\left\{\sigma_1^2 = c\right\}, \qquad \text{(B1)}
\end{aligned}
$$

where the Jacobian of the transformation $\mathcal{J}_{(\cdot)\to(\cdot)}$ adds a factor of $(\alpha^2)^K$ (see Appendix B3) 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 last expression, proving Proposition 3.1. Corollary 3.1 is obtained by integrating out α^2 from Eq. (B1), such that $p(\sigma_0^2, \ldots, \sigma_K^2) = \int p(\alpha^2, \sigma_0^2, \sigma_2^2, \ldots, \sigma_K^2)\, d\alpha^2$. □

B2 Implied prior distribution in R2a case

This proves Proposition 3.2 and Corollary 3.3:

Proof. From the variable transformation expressed in Eq. (20), the joint prior distribution of the corresponding parameters in the restricted model and of the working parameter is derived as follows:

$$
\begin{aligned}
p(\Phi, \Sigma, \alpha^2) &= \mathcal{J}_{(\widetilde{\Phi},\widetilde{\Sigma})\to(\Phi,\Sigma,\alpha^2)}\, p(\widetilde{\Phi}, \widetilde{\Sigma}), \\
&\propto \left(\alpha^2\right)^{\frac{P(P+1)}{2}+K} \left|\widetilde{\Phi}\right|^{-\frac{\nu_0+P+1}{2}} \exp\left\{-\frac{1}{2}\mathrm{tr}\left(t_0 S_0 (\widetilde{\Phi})^{-1}\right)\right\} \\
&\qquad \times \prod_{k=0}^{K} \left(\widetilde{\sigma}_k^2\right)^{-a_0-1} \exp\left\{-\frac{t_0 b_0}{\widetilde{\sigma}_k^2}\right\} \mathbb{1}\left\{\Phi_{[1,1]} + \sigma_0^2 + \sigma_1^2 = c\right\}, \\
&\propto \left(\alpha^2\right)^{-(\nu_0 P + 2a_0(K+1))/2-1} \exp\left\{-\frac{t_0}{2\alpha^2}\left[\mathrm{tr}\left(S_0(\Phi)^{-1}\right) + 2b_0\sum_{k=0}^{K}\frac{1}{\sigma_k^2}\right]\right\} \\
&\qquad \times |\Phi|^{-\frac{\nu_0+P+1}{2}} \prod_{k=0}^{K} \left(\sigma_k^2\right)^{-a_0-1} \mathbb{1}\left\{\Phi_{[1,1]} + \sigma_0^2 + \sigma_1^2 = c\right\}, \qquad \text{(B2)}
\end{aligned}
$$

where $\mathcal{J}_{(\widetilde{\Phi},\widetilde{\Sigma})\to(\Phi,\Sigma,\alpha^2)} \propto (\alpha^2)^{\frac{P(P+1)}{2}+K}$ is the Jacobian of the corresponding transformation (see Appendix B5). The kernel of a scaled inverse chi-squared distribution can be extracted from Eq. (B2), proving Proposition 3.2. Corollary 3.3 is obtained by integrating out α^2 from Eq. (B2), such that $p(\Phi, \Sigma) = \int p(\Phi, \Sigma, \alpha^2) d\alpha^2$. □

B3 Jacobian of the variable transformation in **R1a** case

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 $(\alpha^2)^K$.

B4 Jacobian of the variable transformation in **R1b** case

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 $(\alpha^2)^K$.

B5 Jacobian of the variable transformation in R2a case

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:[23]

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

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:

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

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

23. 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.

Similarly to Appendix B4, 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 $(\alpha^2)^{\frac{P(P+1)}{2}+K}$.

B6 Jacobian of the variable transformation in R2b case

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:

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

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 $(\alpha^2)^{\frac{P(P+1)}{2}+K}$.

C Details of the MCMC sampler

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

Since $p(\widetilde{Y}, \alpha^2 \mid D, \beta, \Phi, \Sigma) \propto p(\widetilde{Y} \mid D, \alpha^2, \beta, \Phi, \Sigma) p(\alpha^2 \mid \Phi, \Sigma)$, the sampling is done by first sampling α^2 from its prior specified in Eq. (22), then drawing Y from its conditional distribution given D, 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. (28)), the conditional distribution of $\widetilde{\beta}$ is:

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

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:

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

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

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

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:

$$\begin{aligned}
\delta_\alpha &= NK + \nu_0 P + 2a_0(K+1),\\
\tau_\alpha &= \sum_{i=1}^{N}(\widetilde{Y}_i - X_i\widehat{\beta})'(\Gamma\Phi\Gamma' + \Sigma)^{-1}(\widetilde{Y}_i - X_i\widehat{\beta}) + \widehat{\beta}'B_0^{-1}\widehat{\beta}\\
&\qquad + t_0\mathrm{tr}\left(S_0(\Phi)^{-1}\right) + 2t_0b_0\sum_{k=0}^{K}\frac{1}{\sigma_k^2},
\end{aligned}$$

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 latent factors 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 latent factors η_i represent the regression

coefficients associated with the matrix of indicators Γ. Therefore, in the expanded model the latent factors are sampled from:

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

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:[24]

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

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 latent factors 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 $p(\widetilde{\Sigma}, \widetilde{\Phi}, \alpha^2 \mid \widetilde{Z}, \widetilde{\eta}, \widetilde{u}_0) = p(\alpha^2 \mid \widetilde{\Sigma}, \widetilde{\Phi})\, p(\widetilde{\Phi} \mid \widetilde{\eta})\, p(\widetilde{\Sigma} \mid \widetilde{Z}, \widetilde{u}_0)$, where $\widetilde{Z} = (\widetilde{Z}_1, \ldots, \widetilde{Z}_N)'$, with $\widetilde{Z}_i = \widetilde{Y}_i - X_i\widetilde{\beta} - \Gamma\widetilde{\eta}_i + \widetilde{u}_{0i}$, for $i = 1, \ldots, N$. Both schemes start by sampling the idiosyncratic variances, 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:

$$\widetilde{\sigma}_0^2 \mid \widetilde{u}_0 \sim \mathcal{G}^{-1}\left(a_0 + \frac{N}{2};\, t_0 b_0 + \frac{1}{2}\sum_{i=1}^{N}(\widetilde{u}_{0i})^2\right),$$
$$\widetilde{\sigma}_k^2 \mid \widetilde{Z} \sim \mathcal{G}^{-1}\left(a_0 + \frac{N}{2};\, t_0 b_0 + \frac{1}{2}\sum_{i=1}^{N}(\widetilde{Z}_{ik})^2\right),$$

for $k = 1, \ldots, K$, where $\widetilde{Z}_{ik}$ denotes the kth element of the vector $\widetilde{Z}_i$, and where $t_0 = 1$ for the two R1* schemes.

The covariance matrix of the latent factors $\widetilde{\Phi}$ is updated in the expanded version of the

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

model from the following inverse-Wishart distribution:

$$\widetilde{\Phi} \mid \widetilde{\eta} \sim \begin{cases} \mathcal{W}^{-1}\Big(\nu_0 + N;\, \widetilde{\eta}'\widetilde{\eta} + \widetilde{S}_0\Big)\,, & \text{for schemes R1a and R1b,} \\ \mathcal{W}^{-1}(\nu_0 + N;\, \widetilde{\eta}'\widetilde{\eta} + t_0 S_0)\,, & \text{for schemes R2a and R2b,} \end{cases}$$

where $\widetilde{S}_0 = \alpha^2 S_0$, using the previous value of the working parameter α^2.

Given the idiosyncratic variances and the covariance matrix of the latent factors in the expanded model, the working parameter is deterministic, i.e., $p(\alpha^2 \mid \widetilde{\Sigma}, \widetilde{\Phi})$ is equal to 1 if α^2 has the expected value given $\widetilde{\Sigma}$ and $\widetilde{\Phi}$, to 0 otherwise. Therefore, 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). For the last two, it is calculated as $\alpha^2 = \left[\widetilde{\Phi}^{\star}_{[1,1]} + \widetilde{\sigma}_0^2 + \widetilde{\sigma}_1^2\right]/c$ (scheme R2a) or as $\alpha^2 = \left[\mathrm{tr}\left(\widetilde{\Phi}\right) + \sum_{k=0}^{K} \widetilde{\sigma}_k^2\right]/c$ (scheme R2b).

References

Altonji, J. G., P. Bharadwaj, and F. Lange. 2012. "Changes in the Characteristics of American Youth: Implications for Adult Outcomes." *Journal of Labor Economics* 30 (4): 783–828. doi:10.1086/666536.

Altonji, J. G., and C. R. Pierret. 2001. "Employer Learning and Statistical Discrimination." *The Quarterly Journal of Economics* 116 (1): 313–350.

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.

Bureau of Labor Statistics, U.S. Department of Labor. 2014. *National Longitudinal Survey of Youth 1979 cohort, 1979-2012 (rounds 1-25).* Produced and distributed by the Center for Human Resource Research, The Ohio State University. Columbus, OH.

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.

Cameron, S. V., and J. J. Heckman. 1998. "Life Cycle Schooling and Dynamic Selection Bias: Models and Evidence for Five Cohorts of American Males." *Journal of Political Economy* 106 (2): 262–333. doi:10.1086/250010.

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.

Deming, D. J. 2017. "The Growing Importance of Social Skills in the Labor Market." *Quarterly Journal of Economics* 132 (4): 1593–1640. doi:10.1093/qje/qjx022.

Dohmen, T., A. Falk, D. Huffman, U. Sunde, J. Schupp, and G. G. Wagner. 2011. "Individual risk attitudes: Measurement, determinants, and behavioral consequences." *Journal of the European Economic Association* 9 (3): 522–550. doi:10.1111/j.1542-4774.2011.01015.x.

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.

Haaijer, R., M. Wedel, M. Vriens, and T. Wansbeek. 1998. "Utility Covariances and Context Effects in Conjoint MNP Models." *Marketing Science* 17 (3): 236–252. doi:10.1287/mksc.17.3.236.

Hansen, K. T., J. J. Heckman, and K. J. Mullen. 2004. "The effect of schooling and ability on achievement test scores." *Journal of Econometrics* 121 (1-2): 39–98. doi:10.1016/j.jeconom.2003.10.011.

Heckman, J. J., J. E. Humphries, and G. Veramendi. 2016. "Dynamic treatment effects." *Journal of Econometrics* 191 (2): 276–292. doi:10.1016/j.jeconom.2015.12.001. arXiv: arXiv:1011.1669v3.

Heckman, J. J., J. Stixrud, and S. Urzua. 2006. "The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior." *Journal of Labor Economics* 24 (3): 411–482. doi:10.1086/504455.

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. doi:10.18637/jss.v014.i03.

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.

Neal, D. A., and W. R. Johnson. 1996. "The Role of Premarket Factors in Black-White Wage Differences." *Journal of Political Economy* 104 (5): 869. doi:10.1086/262045.

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.

Nobile, A. 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.

Speer, J. D. 2017. "The Gender Gap in College Major: Revisiting the Role of Pre-College Factors." *Labour Economics* 44:69–88. doi:10.1016/j.labeco.2016.12.004.

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., and X.-L. Meng. 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 Multinomial Probit Model with Latent Factors

Identification and Interpretation without a Measurement System

Web Appendix

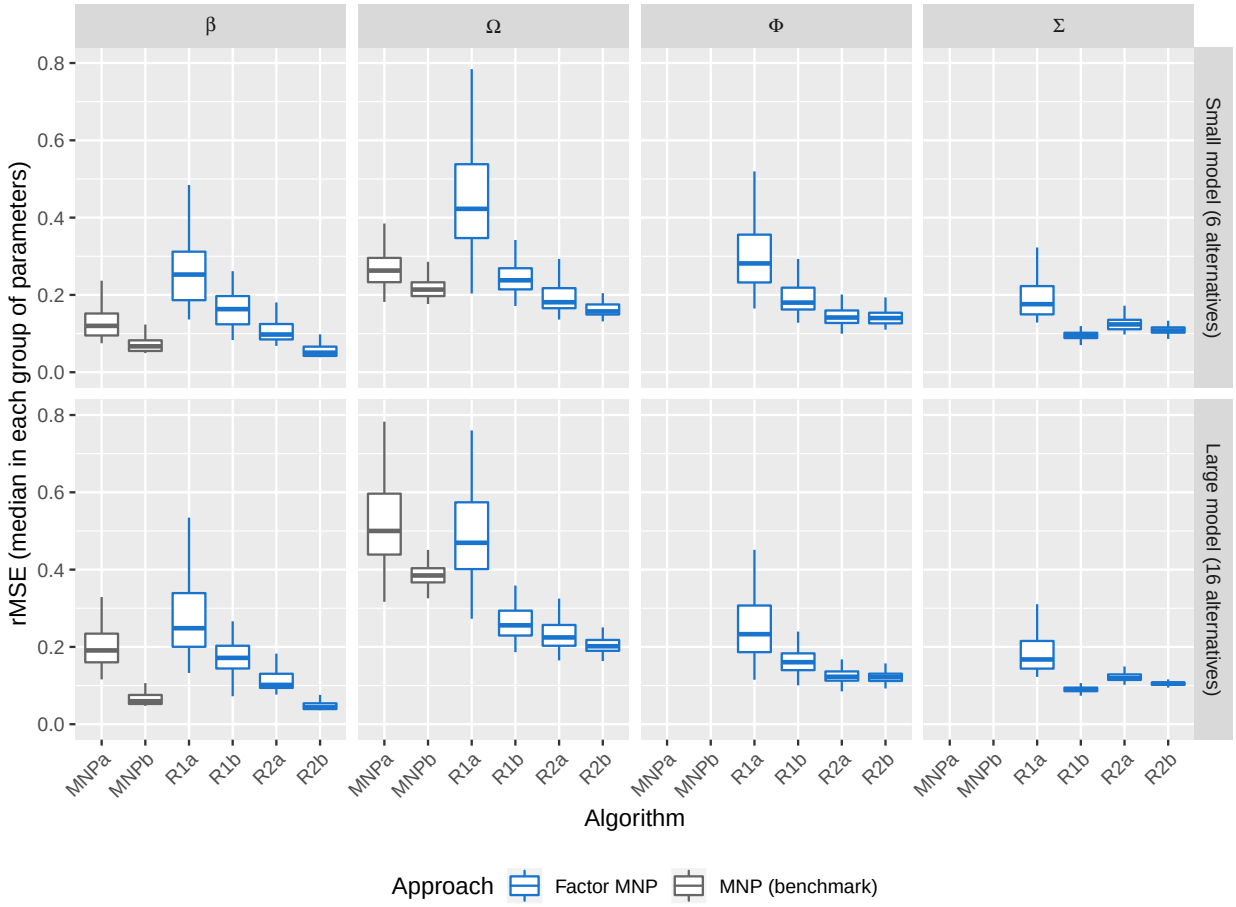
Rémi Piatek
University of Copenhagen
Department of Economics

Miriam Gensowski
University of Copenhagen
Department of Economics
and IZA

Note: *This appendix contains additional results to be made available online, and not to be included in the published version of the paper.*

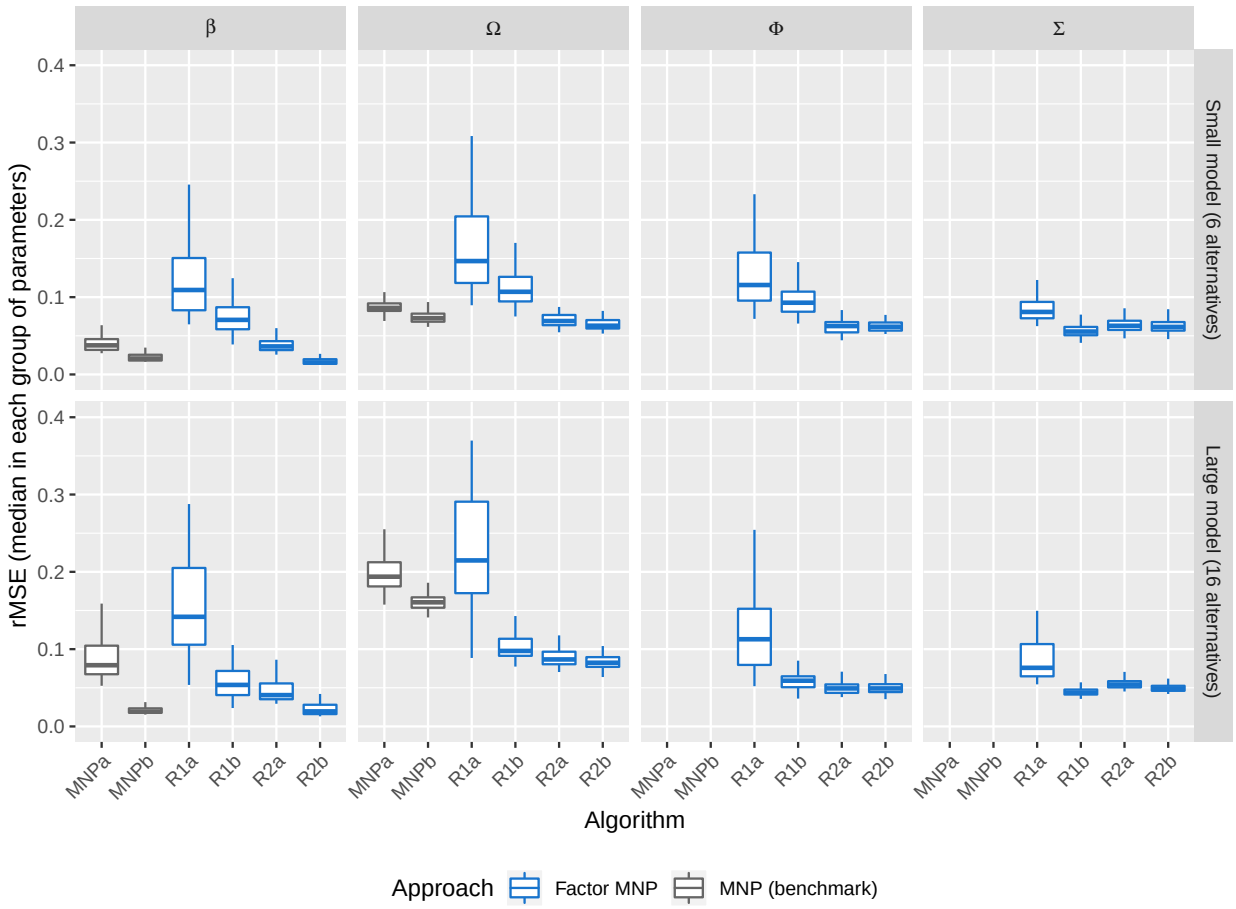
W1 Monte Carlo experiment: Additional results

Figure W1: Simulation study — Root mean square error of model parameters, case with $N = 1,000$ observations.



Notes: Monte Carlo experiment with 100 replications. Box and whiskers plot à la Tukey, where the box shows the 25/50/75th percentiles, and the whiskers 1.5 the inter-quartile range (IQR). Outliers not shown to facilitate the readability of the figure. The statistic of interest is the root mean square error (rMSE) of the parameters, where the median is computed in each group of model parameters for each Monte Carlo replication: Overall covariance matrix of the latent part of the model (Ω), covariance matrix of the latent factors (Φ), idiosyncratic variances of the error terms (Σ). Our approach with the four different restrictions (R1a, R1b, R2a, R2b) is displayed in blue, benchmark MNP model (*without* latent factors, hence no values in panels Φ and Σ) in grey. MNPa: MNP with restriction on first element of the covariance matrix, MNPb: MNP with trace restriction.

Figure W2: Simulation study — Root mean square error of model parameters, case with $N = 10,000$ observations.



Notes: Monte Carlo experiment with 100 replications. Box and whiskers plot à la Tukey, where the box shows the 25/50/75th percentiles, and the whiskers 1.5 the inter-quartile range (IQR). Outliers not shown to facilitate the readability of the figure. The statistic of interest is the root mean square error (rMSE) of the parameters, where the median is computed in each group of model parameters for each Monte Carlo replication: Overall covariance matrix of the latent part of the model (Ω), covariance matrix of the latent factors (Φ), idiosyncratic variances of the error terms (Σ). Our approach with the four different restrictions (R1a, R1b, R2a, R2b) is displayed in blue, benchmark MNP model (*without* latent factors, hence no values in panels Φ and Σ) in grey. MNPa: MNP with restriction on first element of the covariance matrix, MNPb: MNP with trace restriction.

W2 Empirical application: Additional statistics and results

W21 Explanatory variables

The following explanatory variables included as control variables are individual-specific:

- `Minority` is an indicator variable for being described as Black or Hispanic by the NLSY interviewer in 1979.
- `AFQT` is a proxy for cognitive ability.
- `MotherHGC` stands for mother's highest grade completed.
- `ExpectService` is an indicator variable that indicates occupational expectations in 1979 to fall within the Sales/Service occupation group.
- `ExpectBlueCollar` is an indicator variable that indicates occupational expectations in 1979 to fall within a blue-collar occupation.
- `Risk` captures the respondents' willingness to incur risks generally, on a scale from 0 to 10.

Two explanatory variables are alternative-specific:

- `AvgWage` is the average log wage in each education/occupation category.
- `GenderShare` is the share of workers from the respondent's own gender in each education/occupation category.

Table W1: NLSY79 Application — Descriptive statistics.

	Mean	Median	St.dev.	Min	Max
`Married`	0.56	1.00	0.50	0.00	1.00
`Minority`	0.48	0.00	0.50	0.00	1.00
`AFQT`	43.37	39.87	28.79	0.00	100.00
`MotherHGC`	10.91	12.00	3.16	0.00	20.00
`ExpectService`	0.26	0.00	0.44	0.00	1.00
`ExpectBlueCollar`	0.21	0.00	0.40	0.00	1.00
`Risk`	3.97	4.00	3.16	0.00	10.00
`AvgWage`: Low/Business	10.45	10.34	0.33	10.00	10.98
`AvgWage`: Low/Service	9.99	9.94	0.19	9.79	10.50
`AvgWage`: Low/BlueCollar	10.06	10.01	0.21	9.70	10.36
`AvgWage`: High/Business	11.00	10.89	0.31	10.62	11.51
`AvgWage`: High/Service	10.35	10.13	0.32	9.92	10.82
`AvgWage`: High/BlueCollar	10.36	10.38	0.36	9.57	10.72
`GenderShare`: Low/Business	0.50	0.53	0.11	0.36	0.64
`GenderShare`: Low/Service	0.51	0.62	0.18	0.30	0.70
`GenderShare`: Low/BlueCollar	0.49	0.23	0.34	0.12	0.88
`GenderShare`: High/Business	0.50	0.55	0.08	0.40	0.60
`GenderShare`: High/Service	0.51	0.63	0.17	0.31	0.69
`GenderShare`: High/BlueCollar	0.49	0.25	0.31	0.10	0.90
N	5,917				

Notes: Showing the mean, standard deviation (St.dev.), minimum and maximum (Min and Max) of all covariates used in the empirical application in Section 5. For the last two variables, the first part "Low" or "High" denotes education levels, and the second part "Business" corresponds to the Business/Management/STEM occupation group, "Service" includes all service, sales and office workers, and "BlueCollar" all blue-collar occupations.

W22 Additional posterior results

Table W2: NLSY79 Application — Posterior results for the regression coefficients.

	Low Education		High Education		
	Business	Services	Blue Collar	Business	Services
Individual-specific covariates					
`Intercept`	-1.344	0.071	-2.233	-3.060	-1.544
	(0.243)	(0.121)	(0.276)	(0.315)	(0.228)
`Married`	0.182	0.020	0.134	0.307	0.080
	(0.058)	(0.050)	(0.057)	(0.058)	(0.049)
`Minority`	-0.020	-0.035	0.570	0.674	0.488
	(0.081)	(0.064)	(0.091)	(0.085)	(0.078)
`AFQT`	0.012	-0.004	0.019	0.035	0.017
	(0.003)	(0.002)	(0.003)	(0.003)	(0.003)
`MotherHGC`	0.027	0.000	0.065	0.089	0.061
	(0.011)	(0.009)	(0.012)	(0.011)	(0.010)
`ExpectServ`	0.035	0.177	-0.171	-0.206	0.032
	(0.072)	(0.063)	(0.078)	(0.071)	(0.062)
`ExpectBlue`	-0.200	-0.138	-0.402	-0.695	-0.520
	(0.079)	(0.066)	(0.085)	(0.091)	(0.079)
`Risk`	0.016	0.006	-0.003	0.023	0.014
	(0.008)	(0.007)	(0.008)	(0.008)	(0.007)
Alternative-specific covariates					
`AvgWage`			0.119		
			(0.069)		
`GenderShare`			1.071		
			(0.077)		

Notes: Results from the empirical application in Section 5 showing posterior means and standard deviations (in parentheses) of the regression coefficients. Baseline level of the alternatives is "Low Education/Blue Collar." `AFQT` is a proxy for cognitive ability, `MotherHGC` stands for mother's highest grade completed, `ExpectService` is an indicator variable that indicates occupational expectations in 1979 to fall within the Sales/Service occupation group, and `ExpectBlue` the corresponding indicator for expecting to go into a blue collar occupation. `Risk` captures the respondents' willingness to incur risks generally, ranging from 0 to 100.

Table W3: NLSY79 Application — Posterior results for the idiosyncratic variances.

	Low Education			High Education		
	Blue Collar	Business	Services	Blue Collar	Business	Services
σ_j^2	0.212	0.286	0.152	0.259	0.181	0.136
	(0.090)	(0.154)	(0.090)	(0.128)	(0.110)	(0.071)

Notes: Results from the empirical application in Section 5. Showing posterior means and standard deviations (in parentheses) of the idiosyncratic variances. Baseline level of the alternatives is "Low Education/Blue Collar."