A Bayesian Nonparametric Approach to Factor Analysis

Rémi Piatek*
University of Copenhagen
remi.piatek@econ.ku.dk

Omiros Papaspiliopoulos
ICREA–UPF
omiros.papaspiliopoulos@upf.edu

April 18, 2017

Abstract

This paper introduces a new approach to the identification and inference of non-Gaussian factor models based on Bayesian nonparametric methods. It relaxes the usual normality assumption on the latent factors, widely used in practice, which is too restrictive in many cases and likely to distort the results. Our approach, on the contrary, does not impose any particular assumptions on the shape of the distribution of the factors, but still secures the basic requirements for the identification of the model. We design a new sampling scheme based on marginal data augmentation for the inference of mixtures of normals with moment restrictions. This approach is then augmented by the use of a retrospective sampler to allow the inference of a constrained Dirichlet process mixture model for the distribution of the latent factors. We carry out a simulation study to illustrate the methodology and demonstrate its benefits. Our sampler is very efficient in recovering the distribution of the factors, and only generates models that fulfill the identification requirements. A real data example is provided to illustrate the applicability of the approach.

JEL Classification: C11; C38; C63.

Keywords: Factor Models; Model Identification; Bayesian Nonparametric Methods; Dirichlet Process Hierarchical Models; Retrospective Sampler; Marginal Data Augmentation.

*Corresponding author: Department of Economics, University of Copenhagen, Øster Farimagsgade 5, Building 26, DK–1353 Copenhagen K, Denmark. Phone: (+45) 35 32 30 35. Rémi Piatek's research is 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. The methodology introduced in this paper will be released as an extension to the `R` package `BayesFM` available on `CRAN` at https://cran.r-project.org/package=BayesFM.

Contents

1 Introduction

[OM: Early on we need a more succinct description of the main novelties in this paper relative to what has been done in econometrics and machine learning. We can talk about this]

Factor analysis has grown as a very popular and powerful tool in many fields of research, and particularly in the social sciences, where it is often used to aggregate large sets of variables into smaller sets of meaningful factors. A myriad of examples relying on this data reduction strategy can be found in the empirical literature, ranging from the extraction of latent factors underlying macroeconomic indicators to explain monetary policies or business cycles (Bernanke et al., 2005; Forni and Gambetti, 2010), to the measurement of personality traits and cognitive abilities and their impact on economic outcomes (see, e.g., Carneiro et al., 2003; Hansen et al., 2004; Heckman et al., 2006; Conti et al., 2014).

One of the main challenges inherent to the inference of these models is identification. To make inference feasible and produce meaningful results, identification assumptions are needed, often in the form of parameter restrictions and distributional assumptions. Techniques for dealing with such issues were developed as early as Anderson and Rubin (1956). Within the social sciences, most of the papers published up to date assume the factors to be Gaussian, although within the Machine Learning community, non-Gaussian factor analysis is popular. The Gaussian assumption is convenient and has a natural interpretation for many analysts. However, it may have little justification empirically, and the misspecification it can induce is likely to contaminate the inference of the remaining parameters of the model. In fact, working with non-Gaussian distribution for the factors removes some of the identifiability issues that affect traditional factor analysis.

This paper offers a more flexible approach to factor analysis that relaxes the Gaussian assumption on the latent factors. Relative to the existing literature, we offer a modeling framework that allows for dependence across factors, assumes a flexible distribution based on mixtures of Gaussians, and permits the identification of the latent factors. Dependence across factors and identifiability are key requirements in the applications we are interested in, as they allow to unravel rich latent structures where unobserved traits can be correlated, and interpretation is facilitated—see Almlund et al. (2011) for a discussion on this topic in personality economics. In the econometric literature, estimation methods relying on finite mixtures of normals are commonly used (Hansen et al., 2004; Cunha and Heckman, 2008; Cunha et al., 2010). Mixtures of normals provide an approximation to the unknown distribution of the latent factors that can otherwise be non-parametrically identified through appropriate parameter restrictions. These approaches therefore guarantee identification and ensure interpretability. However, another type of misspecification can emerge when the number of mixture components selected is not appropriate to provide a good fit to the data, which raises additional complications.

Bayesian nonparametric (BNP) methods have been introduced to avoid having to decide *a priori* on the number of components, but instead learn it from the data. See

Antoniak (1974), Quintana and Müller (2004), and Paisley and Carin (2009) for Dirichlet processes in general, Neal (2000) and Papaspiliopoulos and Roberts (2008) for state-of-the-art computational methods for estimating such models, and Yang et al. (2010) and Ghosh and Dunson (2009) for previous use of these models for factor analysis. Effectively, these approaches specify an infinite mixture of Gaussians with specific prior distribution on the weights, and the number of active components can grow with the size of the data. Typically, these procedures do not provide any guarantee of the formal identification of the model. This is not a problem when factor models are used to perform variance decomposition, model shrinkage, low-rank approximation or to do forecasting. However, the lack of identification becomes a major obstacle when the inference of the structural part of the model is of main interest—e.g., if factor loadings need to be identified to make inference on elasticities or marginal effects. A relevant related and active literature is that of independent factor analysis popular in Machine Learning, see for example Attias (1999). In that framework, identifiability is of real concern since the latent factors are used for signal reconstruction. A key observation is that identifiability can be partially resolved by working with certain non-Gaussian distributions for the latent factors.

The goal of the present paper is to develop a richly parameterized and flexible distribution for the latent factors, which allows for dependence among factors while ensuring their identifiability. We specify the distribution of the latent factors as an affine transformation of a Dirichlet process that has marginal means equal to 0 and marginal variances equal to 1. [OM: return to this - not true in the infinite case] We achieve this by appropriately transforming the parameters of the mixture components. We develop a new approach for the inference of constrained mixtures of Gaussians that relies on Marginal Data Augmentation (MDA) methods (Meng and van Dyk, 1999; van Dyk and Meng, 2001; van Dyk, 2010) and combine those with the retrospective sampling algorithm of Papaspiliopoulos and Roberts (2008) to deal with the infinite number of components. MDA methods proceed by expanding the original constrained model, introducing extra parameters that cannot be identified from the data, but facilitate sampling and make inference more efficient in terms of convergence and mixing. An interesting by-product is that the model expansion can be tailored to safeguard the identification of the factor model. In the case of a Dirichlet process mixture model, where the number of mixture components is free to grow infinitely to accommodate the data, the implementation of MDA methods is not straightforward. In this article, we resort to truncations of the Dirichlet process, as in Ishwaran and James (2002). We also explore the unbounded case with the retrospective sampling ideas in Papaspiliopoulos and Roberts (2008) and Yau et al. (2011) to infer the infinite-dimensional mixture model. As shown in the aforementioned articles, avoiding the truncation typically also leads to more efficient computational algorithms. [OM: revisit the whole thing]

With this representation based on a mixture of normals, our model can be reformulated as a mixture of factor analyzers (McLachlan and Peel, 2000; McLachlan et al., 2003; Fokoué and Titterington, 2003). Despite the analogy of the two approaches, there are fundamental

differences. Mixtures of factor analyzers assume Gaussian factors and mix the structure parameters of the model (factor loadings, intercepts and error term variances), while our approach assumes that those are fixed across mixture components and rather mixes the moments of the distribution of the latent factors. Not only does this change completely the interpretation of the model, it also results in different statistical properties of the two approaches.

We conduct a simulation study to investigate the performance of our sampler, using a synthetic data set generated from a two-factor model with a non-standard distribution for the latent factors. The results are very promising. They show that our MCMC sampling scheme succeeds in retrieving the true underlying distribution of the latent factors, without any *a priori* assumptions on the shape of the distribution. Most importantly, it does so in generating identified models only. Sampling turns out to be highly efficient thanks to the MDA procedure. The mixing of the Markov chains is indeed very good compared to what can usually be achieved in latent variable models, where convergence can be prohibitively slow and mixing bad.

The baseline factor model used throughout this paper is presented in section 2, and its identification is also discussed. We briefly outline the parametric identification of the structural part of the model, then spend some time on the nonparametric identification of the distribution of the latent factors, which is our main focus. section 3 introduces the Marginal Data Augmentation sampling scheme for a finite mixture of normal distributions. section 4.1 carries out a simulation study that shows the efficiency of the approach, and section 4.2 applies it to a real data example to illustrate its relevance in practice. section 5 concludes on a few final remarks. The algorithms developed in this paper will be released as an extension to the `R` package `BayesFM`,[1] to allow applied researchers to implement these methods in a user-friendly manner.

2 Specification and identification of the factor model

2.1 General model structure

The generic structure of the latent factor model we are considering is as follows. There are Q manifest variables Y_i and P latent factors θ_i ($P \ll Q$), for $i = 1, \ldots, N$, following a linear relationship through a matrix of factor loadings Λ and a vector of intercept terms δ:

$$\begin{aligned} \underset{(Q\times 1)}{Y_i} &= \underset{(Q\times 1)}{\delta} + \underset{(Q\times P)}{\Lambda}\underset{(P\times 1)}{\theta_i} + \underset{(Q\times 1)}{\varepsilon_i}\,, \\ \varepsilon_i &\sim \mathcal{N}(0;\, \Sigma)\,, \qquad \Sigma = \operatorname{diag}(\sigma_1^2, \ldots, \sigma_Q^2), \end{aligned} \tag{1}$$

[1] available on `CRAN` at https://cran.r-project.org/package=BayesFM.

where the error terms ε_i are assumed to be Gaussian for the sake of simplicity.[2] The independence of the error terms is standard in factor analysis, and implies that the factors are the only source of correlation between the observed variables. The statistical model requires a specification for the distribution of the latent factors. A default option in the literature is that of a Gaussian distribution. In this paper, we relax this assumption by allowing a nonparametric specification of this distribution. Bayesian inference for this factor model will also require priors on δ, Λ, Σ, and typically other hyperparameters as well.

The model as stated in eq. (1) is not identified. Some identifiability issues arise because of the specification of the structural part of the model (section 2.2), while others are related to the distributional assumptions on the latent factors (section 2.3). The lack of identifiability becomes a major obstacle when the inference of the structural part of the model is of main interest, for instance to identify and make inference on policy-relevant parameters—e.g., elasticities, marginal effects or latent traits. The following sections provide an overview of the two sources of identification issues that lead to our proposal for an identifiable nonparametric latent factor model.

2.2 Identification of the structural part of the model

The intercept terms δ and the factor loading matrix Λ can be identified using appropriate parameter restrictions. This can be done independently of the distribution assumed on the latent factors, but we will also see that specific distributional assumptions can allow to relax some of these restrictions.

If the distribution of the latent factors belongs to a location family, such as the Gaussian, or a mixture of distributions in a location family, such as mixture of Gaussians, with location parameter(s) to be estimated from the data, then δ is not identifiable. Indeed, the distribution of Y_i remains the same by adding an arbitrary constant to δ and subtracting appropriate constant(s) from the location parameter(s). This lack of identifiability can be tackled by fixing the location of the factors, e.g., by fixing the mean of the factor distribution to 0. This constraint is straightforward to impose in the Gaussian case, but not as trivial in the nonparametric case. In this article, we propose a distribution for the factors that fixes their location.

The second identification problem affects the factor loadings. The bilinear form $\Lambda\theta_i$ implies that the latent factors can only be identified up to a scale transformation, since the distribution of Y_i remains unaltered if the factors are multiplied by a nonsingular scaling matrix, and the factor loading matrix by the inverse of this matrix. This can be seen from the expression of the overall covariance matrix of the manifest variables, which can be expressed as $\Lambda\Phi\Lambda' + \Sigma = (\Lambda R^{-1})(R\Phi R')(\Lambda R^{-1})' + \Sigma$, for any nonsingular $(P \times P)$-matrix R, where $\Phi \equiv \mathrm{V}(\theta_i)$ denotes the covariance matrix of the latent factors.

[2] The normality of the error terms could be relaxed in a similar way to the latent factors. However, we stick to the standard Gaussian assumption in this paper for simplicity, and because the main focus is on the distribution of the latent factors.

This indeterminacy, commonly referred to as the *rotation problem*, is well known since the seminal work of Thurstone (1934), later formalized by Reiersøl (1950), Koopmans and Reiersøl (1950), and Anderson and Rubin (1956).

This lack of identifiability has been addressed in the literature by assuming that the latent factors are uncorrelated and have unit variances, such that $\Phi = I_P$. This requirement has made the standard Gaussian a distribution of choice in factor analysis. However, this assumption does not completely solve the indeterminacy problem, as the system still remains unchanged if R is specified as an orthogonal matrix. To rule out these cases, Anderson and Rubin (1956, p. 121) propose to use a lower triangular structure for the upper part of Λ. This structure has then become popular in factor analysis, see, e.g., Geweke and Zhou (1996), Aguilar and West (2000), Lopes and West (2004), and Frühwirth-Schnatter and Lopes (2010).

One last identifiability issue needs to be taken care of. It arises because the sign of the latent factors and of the corresponding columns of the loading matrix can be flipped simultaneously without affecting the distribution of Y_i. This property of the model implies that without further constraints on Λ (or on the factors) the sign of the correlation between the factors is not identifiable. In our work, we fix the scale of the factors and deal with the sign issue by making assumptions on the sign of certain entries of the loading matrix. Computationally, we work with the sign-unconstrained model and enforce the constrains at a post-processing stage by appropriate transformation of the MCMC output, as in, e.g., Frühwirth-Schnatter and Lopes (2010) and Conti et al. (2014), see section 3.

Alternatively, the scales and the signs of the latent factors can be set by constraining one loading in each column of Λ instead of constraining the diagonal elements of Φ. This approach has been popular in the econometrics literature, as it allows to *anchor* the latent factors in real measurements, thus facilitating interpretation (for example Cunha and Heckman, 2008; Cunha et al., 2010, anchor the factors in earnings outcomes). Nevertheless, constraining some factor loadings can be too restrictive in some frameworks. For example, when a stochastic search is carried out to determine the number of latent factors and the structure of the factor loading matrix in terms of zero and nonzero elements, it is not possible to fix any of the loadings *a priori*. These approaches are becoming increasingly popular in the literature, see, among others, Lucas et al. (2006), Carvalho et al. (2008), Frühwirth-Schnatter and Lopes (2010), Bhattacharya and Dunson (2011), and Conti et al. (2014). In the present paper, we rely on identifying criteria that fix the variances of the factors rather than some of the factor loadings.

When working with correlated factors, the lower triangular block structure of Λ no longer safeguards identification. Indeed, pre-multiplying the latent factors by a nonsingular lower-triangular matrix R and post-multiplying Λ by the inverse of R results in a model that is observationally equivalent to the original one, since ΛR^{-1} also has a lower-triangular block structure. Therefore, the release of the constraints on the lower-diagonal elements of Φ used in the uncorrelated case needs to be compensated by additional constraints on the factor loading matrix in the correlated case. This can be done by specifying a diagonal ma-

trix for the upper part of Λ, such that $\Lambda' = \begin{pmatrix} D_{\Lambda_1} & \Lambda_2' \end{pmatrix}$, with $D_{\Lambda_1} = \text{diag}(\lambda_{11}, \ldots, \lambda_{PP})$, and Λ_2 is a full matrix that may contain additional zero elements. In this specification, the first P manifest variables each load on a single latent factor, and are sometimes called *dedicated* measurements in the literature (Conti et al., 2014; Williams, 2015). Similarly to the uncorrelated case, the scale of the factors is set by either assuming that $D_{\Lambda_1} = I_P$, or that $\Phi_{pp} = 1$ and $\lambda_{pp} > 0$, for $p = 1, \ldots, P$.

2.3 Nonparametric identification of the distribution of the latent factors

The restrictions derived in section 2.2 allow to achieve identification of the structural part of the model, i.e., δ and Λ, and also of the covariance matrix of the latent factors Φ. Notably, these assumptions do not depend on the distributional assumptions made on the latent factors. Importantly, they only secure the identification of the covariance matrix of the factors, and therefore do not guarantee that the whole distribution of the factors is identified if we depart from the Gaussian case.

These assumptions might be over-restrictive in the nonparametric case, depending on the approach adopted. For example, working with non-Gaussian latent factors can remove some identifiability problems when the latent factors follow a mixture of Gaussians with diagonal covariance matrix for each component but different from the identity. This property has propelled the so-called independent component analysis and independent factor analysis, popular within Machine Learning, see, e.g., Attias (1999).

On the other hand, some nonparametric approaches might require additional restrictions to fully identify the distribution of the factors nonparametrically. In this paper, we rely on the identification strategy developed in Cunha et al. (2010). Their nonparametric approach requires mild assumptions on the latent factors, and only minor additional restrictions on the factor loading matrix, since two dedicated manifest variables are needed for each factor instead of one in the previous section.[3]

With two dedicated manifest variables in hand for each latent factor, such that $\Lambda' = \begin{pmatrix} D_{\Lambda_1} & D_{\Lambda_2} & \Lambda_3' \end{pmatrix}$,[4] the proof for nonparametric identification of the factor distribution follows from Cunha et al. (2010). Assuming nonzero diagonal elements in D_{Λ_1} and D_{Λ_2}, the first $2P$ equations can be rewritten as

$$\begin{aligned} W_1 &= \theta + \omega_1, \\ W_2 &= \theta + \omega_2, \end{aligned} \tag{2}$$

with

$$\begin{aligned} W_1 &= D_{\Lambda_1}^{-1} \left(Y_{1:P} - \delta_{1:P} \right), & \omega_1 &= D_{\Lambda_1}^{-1} \varepsilon_{1:P}, \\ W_2 &= D_{\Lambda_2}^{-1} \left(Y_{(P+1):(2P)} - \delta_{(P+1):(2P)} \right), & \omega_2 &= D_{\Lambda_2}^{-1} \varepsilon_{(P+1):(2P)}, \end{aligned}$$

[3] In most cases, the assumption of two dedicated measurements per factor is not restrictive in practice, since numerous indicators are usually available to measure the latent factors.

[4] Similarly to section 2.2, D_{Λ_1} and D_{Λ_2} are diagonal matrices, Λ_3 is a full matrix.

where the subscripts denote the elements of the corresponding subvectors (e.g., $Y_{1:P}$ contains the first P elements of the vector Y). The expression of the subsystem corresponding to the dedicated measurements in eq. (2) is particularly convenient, as it allows to directly use the first theorem of Cunha et al. (2010, Theorem 1, p. 893) to prove the nonparametric identification of the distribution of the factors, after having secured the identification of the intercept terms and the factor loadings, as explained in the previous section. This theorem states that if W_1, W_2, θ, ω_1 and ω_2 are random vectors taking values in $\mathbb{R}^P$ and related through the equations in eq. (2), then the factor distribution is nonparametrically identified and can be expressed in terms of observable quantities, provided that $\mathrm{E}\Big(\omega_1 \mid \widetilde{\theta}, \omega_2\Big) = 0$ and ω_2 is independent from θ. The last two conditions are automatically fulfilled, since we assume the error terms to be independently normally distributed.

2.4 Identifiable Bayesian nonparametric correlated factor models

We build a model for the latent factors that is sufficiently constrained in its location and scale to facilitate identifiability of the structural part of the model. The model is constructed as an affine transformation of an auxiliary process, which is modeled as a Dirichlet process Gaussian mixture model and is described below. Therefore, our approach is a combination of Bayesian nonparametrics and econometric modeling in order to ensure both a flexible model for the latent factors and identifiability of the structural part of the model. It turns out that an insightful perspective on our model is as a Gaussian mixture model where the number of components can be learned from the data automatically, and where the mixture parameters are constrained to ensure identifiability of the structural part of the factor model. The induced constraints lead to a complicated posterior distribution, but we propose marginal data augmentation methods in section 3 to sample from it very efficiently.

2.4.1 Modeling the distribution of the factors

In the rest of the paper we will follow a notational convention. The intercept, factor loadings and latent factors that appear in the final formulation of the factor model will be denoted by δ, Λ and θ_i, respectively, whereas transformations thereof by $\widetilde{\delta}$, $\widetilde{\Lambda}$ and $\widetilde{\theta}_i$. These transformations might be used as intermediate variables in the construction of the final model, e.g., an intermediate $\widetilde{\theta}_i$ is used to define a model for factors θ_i with constraints on their location and scale, or as working parameters to construct better MCMC algorithms to estimate the factor model, e.g., $\widetilde{\delta}$, $\widetilde{\Lambda}$, or both, implying that $\widetilde{\theta}_i$ is also used to define working parameters for efficient MCMC. Below we explain the precise ways that these transformations relate to eq. (1).

The factor model is an affine transformation of an auxiliary process that is modelled

as a Dirichlet process Gaussian mixture model, as we describe below:

$$\begin{aligned}
\widetilde{\theta}_i \mid \widetilde{\mu}_{G_i}, \widetilde{\Phi}_{G_i} &\sim \mathcal{N}\left(\widetilde{\mu}_{G_i};\, \widetilde{\Phi}_{G_i}\right), \\
G_i \mid p &\sim \sum_{k=1}^{K} p_k \delta_k(G_i), \\
\widetilde{\mu}_k \mid \widetilde{\Phi}_k, A_0 &\sim \mathcal{N}\left(0;\, A_0 \widetilde{\Phi}_k\right), \qquad (3) \\
\widetilde{\Phi}_k \mid \nu_0, s_0 &\sim \mathcal{IW}(\nu_0;\, s_0 I_P), \qquad (4) \\
p_1 = V_1, \quad p_k &= V_k \prod_{j=1}^{k-1} (1 - V_j), \qquad (5) \\
V_k &\sim \mathcal{B}eta(1; \alpha), \\
1 < k &\leq K.
\end{aligned}$$

The parameters that define the Gaussian distribution at the top level are denoted by $\widetilde{\vartheta}_k = \{\widetilde{\mu}_k, \widetilde{\Phi}_k\}$ and are collected in the set $\widetilde{\vartheta} = \{\widetilde{\vartheta}_1, \widetilde{\vartheta}_2, \ldots\}$. These and the random variables $V = (V_1, V_2, \ldots)$, are assumed to be independent of each other. The Dirac delta function centered at k is denoted $\delta_k(\cdot)$. Hence, when $K = \infty$, lines 2 to 6 in the above hierarchy define a Dirichlet process model for $\widetilde{\vartheta}$ with base distribution normal-inverse-Wishart (defined in lines 3-4) parameterized by $\{A_0, \nu_0, S_0\}$. When $K < \infty$, V_K is set to 1 to ensure that the mixture weights sum up to 1, and the resultant model is a truncated Dirichlet process for $\widetilde{\vartheta}$ (Ishwaran and James, 2001, 2002). In either case, we adopt the stick-breaking representation for the Dirichlet process (Sethuraman, 1994), as described in lines 5-6 in the model, and we explicitly augment the model with latent variables $G = (G_1, \ldots, G_N)$ for the mixture group memberships. Marginalising over these membership variables, we obtain a (potentially infinite) mixture of Gaussians for the distribution of the factors:

$$\widetilde{\theta}_i \sim \sum_{k=1}^{K} p_k \mathcal{N}\left(\widetilde{\mu}_k;\, \widetilde{\Phi}_k\right),$$

with corresponding moments:

$$\begin{aligned}
\mathrm{E}\left(\widetilde{\theta}_i\right) &= \sum_{k=1}^{K} p_k \widetilde{\mu}_k, \\
\mathrm{V}\left(\widetilde{\theta}_i\right) &= \sum_{k=1}^{K} p_k \left((\widetilde{\mu}_k - \widetilde{\mu})(\widetilde{\mu}_k - \widetilde{\mu})' + \widetilde{\Phi}_k\right).
\end{aligned}$$

2.4.2 Constrained version of the model

Relying on the distribution specified in section 2.4.1 for $\widetilde{\theta}_i$, we propose the following identifiable nonparametric factor model:

$$\begin{aligned} Y_i &= \delta + \Lambda\,\theta_i + \varepsilon_i, \\ \theta_i &= \widetilde{D}^{-\frac{1}{2}}(\widetilde{\theta}_i - \widetilde{\mu}). \end{aligned} \tag{6}$$

We choose $\widetilde{\mu}$ and $\widetilde{D}$ so as to constrain the location and scale of the latent factors. We treat the finite ($K < \infty$) and infinite ($K = \infty$) mixture cases separately, although both are based on the following construction:

$$\widetilde{\mu} = \sum_{k=1}^{K} \beta_k \widetilde{\mu}_k, \tag{7}$$

$$\widetilde{\Phi} = \sum_{k=1}^{K} \beta_k \left((\widetilde{\mu}_k - \widetilde{\mu})(\widetilde{\mu}_k - \widetilde{\mu})' + \widetilde{\Phi}_k \right), \qquad \widetilde{D} \equiv \operatorname{diag}(\widetilde{\Phi}_{11}, \ldots, \widetilde{\Phi}_{PP}), \tag{8}$$

where the weights β_k are chosen in different ways for the finite and infinite mixture models. The structure of the factor loading matrix, in terms of zero restrictions, is not affected by the transformation because the matrix $\widetilde{D}$ used to rescale the latent factors is diagonal. This is particularly important, as zero restrictions on Λ are required for identification—in particular, two dedicated measurements are required for each latent factor, resulting in zero restrictions in the corresponding rows of Λ, see section 2.2.[5,6] Since the Gaussian is a location-scale family, it follows that an equivalent way to understand the proposed latent factor model is as a Gaussian mixture with linearly constrained parameters:

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

$$\mu_k = \widetilde{D}^{-1/2}(\widetilde{\mu}_k - \widetilde{\mu}) \tag{9}$$

$$\Phi_k = \widetilde{D}^{-1/2}\widetilde{\Phi}_k\widetilde{D}^{-1/2}, \tag{10}$$

which by construction implies that the following constraints are fulfilled: [REMI: I added the following definition that was missing and I think is very important, as it shows how

[5] If sign restrictions are imposed on Λ for identification, these restrictions also remain unaffected by the expansion, since the diagonal elements of $\widetilde{D}$ are all positive.

[6] This last remark explains why we only use the variances of the latent factors to expand the model, and not the Cholesky decomposition of the covariance matrix $\widetilde{\Phi}$. While this latter approach would be appropriate in a model with uncorrelated factors where a lower triangular block structure is used for the factor loading matrix, in our case with dedicated measurements the zero restrictions also required in the lower part of the triangle would not be preserved by the parameter expansion.

we constrain the location and scale of the factors]

$$\mu \equiv \sum_{k=1}^{K} \beta_k \mu_k = 0_P, \tag{11}$$

$$\Phi \equiv \sum_{k=1}^{K} \beta_k \left(\mu_k \mu_k' + \Phi_k\right), \qquad D \equiv \operatorname{diag}(\Phi_{11}, \ldots, \Phi_{PP}) = I_P. \tag{12}$$

The prior on $\widetilde{\vartheta}$ specified in eqs. (3) and (4) implies one for the corresponding constrained parameters $\vartheta = \{\vartheta_1, \ldots, \vartheta_K\}$, where $\vartheta_k = \{\mu_k, \Phi_k\}$. The form of this induced density is given in proposition 2 in the Appendix, and specifically in eq. (A1). The density does not belong in a known family and looks rather cumbersome. Fortunately, this density is not required in the sampling scheme, as the marginal data augmentation procedure we will use mainly relies on the expanded version of the model, which is easier to sample from. This should not, however, make us forget to investigate how the prior induced in the identified model is shaped, to make sure we do not work with an odd prior. To do this, it is straightforward to simulate the prior rather than trying to work out its kernel analytically (see section 4.1.2).

The Bayesian formulation of the factor model is complemented by priors on δ, Λ and Σ. The concentration parameter α has a major impact on the estimated number of components in the infinite mixture model, so we prefer to learn it from the data. Therefore, the general structure of the prior distribution on the hyperparameters is

$$\delta \mid c_0 \sim \mathcal{N}_Q(0_Q;\, c_0 I_Q)\,, \tag{13}$$

$$\lambda_q \mid d_0 \sim \mathcal{N}_K(0_K;\, d_0 I_K)\,, \tag{14}$$

$$\sigma_q^2 \mid a_0, b_0 \sim \mathcal{IG}(a_0;\, b_0)\,, \tag{15}$$

$$\alpha \mid g_0, s_0 \sim \mathcal{G}(g_0;\, h_0)\,, \tag{16}$$

for $q = 1, \ldots, Q$, where λ_q denotes the factor loadings on the qth row of Λ.

[OM: Remi I've mentioned you should update alpha too. I have a comment also later on in the article.][REMI: Done.]

2.4.3 Finite mixture model

In the finite mixture model where $K < \infty$ we simply take

$$\widetilde{\mu} \equiv \mathrm{E}\left(\widetilde{\theta}_i\right) = \sum_{k=1}^{K} p_k \widetilde{\mu}_k,$$

$$\widetilde{\Phi} \equiv \mathrm{V}\left(\widetilde{\theta}_i\right) = \sum_{k=1}^{K} p_k \left((\widetilde{\mu}_k - \widetilde{\mu})(\widetilde{\mu}_k - \widetilde{\mu})' + \widetilde{\Phi}_k\right), \qquad \widetilde{D} \equiv \operatorname{diag}\left(\widetilde{\Phi}_{11}, \ldots, \widetilde{\Phi}_{PP}\right).$$

In terms of the generic model structure in eqs. (7) and (8), we take $\beta_k = p_k$. Therefore, eqs. (11) and (12) are by construction equivalent to $\mathrm{E}(\theta_i) = 0_P$ and $\mathrm{diag}\,(\mathrm{V}(\theta_i)) = \iota_P$.

2.4.4 Infinite mixture model

We could repeat the above construction for $K = \infty$, but each component parameter would require an infinite summation in order to be determined, which makes the resulting model computationally intractable. Instead, we use the ingredients of the retrospective sampling methodology of Papaspiliopoulos and Roberts (2008) to define $\widetilde{\mu}$ and $\widetilde{D}$ required in eqs. (6) to (8). The construction now also involves the allocation variables G_i.

In the mixture of Dirichlet processes, the number of mixture components K is nominally infinite, but in practice only a *finite* number of observations N is available and can be allocated to the mixture groups. Therefore, only a finite number of mixture groups will contain observations, the remaining ones being empty mixture components. We introduce some notation, following Papaspiliopoulos and Roberts (2008), and divide the set of mixture component indices ($\mathcal{I}$) into two distinct groups, the group of non-empty mixture components ("alive" components $\mathcal{I}^{(\mathrm{al})}$), and the group of "dead" components ($\mathcal{I}^{(\mathrm{d})}$):

$$\begin{aligned}\mathcal{I} &= \{1, 2, \ldots\},\\ \mathcal{I}^{(\mathrm{al})} &= \{k \in \mathcal{I} : N_k > 0\},\\ \mathcal{I}^{(\mathrm{d})} &= \{k \in \mathcal{I} : N_k = 0\},\end{aligned}$$

where $N_k = \sum_{i=1}^{N} \mathbf{1}\{G_i = k\}$, for $k = 1, 2, \ldots$, such that $\mathcal{I} = \mathcal{I}^{(\mathrm{al})} \cup \mathcal{I}^{(\mathrm{d})}$. Then, we take

$$\begin{aligned}\widetilde{\mu} &\equiv \sum_{k \in \mathcal{I}^{(\mathrm{al})}} w_k \widetilde{\mu}_k,\\ \widetilde{\Phi} &\equiv \sum_{k \in \mathcal{I}^{(\mathrm{al})}} w_k \left((\widetilde{\mu}_k - \widetilde{\mu}^{(\mathrm{al})})(\widetilde{\mu}_k - \widetilde{\mu}^{(\mathrm{al})})' + \widetilde{\Phi}_k \right), \qquad \widetilde{D} \equiv \mathrm{diag}(\widetilde{\Phi}^{(\mathrm{al})}_{11}, \ldots, \widetilde{\Phi}^{(\mathrm{al})}_{PP}),\end{aligned}$$

where $w_k = N_k/N$ denotes the observed frequency of an individual being allocated to mixture component k, where by construction $w_k > 0$ for $k \in \mathcal{I}^{(\mathrm{al})}$, $w_k = 0$ for all $k \in \mathcal{I}^{(\mathrm{d})}$ and $\sum_{k \in \mathcal{I}^{(\mathrm{al})}} w_k = 1$. Therefore, in the notation of eq. (7) and eq. (8) we take $\beta_k = w_k$, which depends on the configuration of the allocation variables. This construction does not collapse to the one for the finite mixture when $K < \infty$. It does not fix the location and scale of the factors by setting their first two moments to fixed values like in the finite case, but rather through the counterparts of these moments constructed in eqs. (11) and (12) that use the observed mixture frequencies. These alternative constraints provides identifiability to the structural part of the model.

[OM: Remi link this better with the results].

[REMI: Could we justify this better? Intuitively it seems to make sense to think of an auxiliary mixture constructed with the observed frequencies as weights, and to impose the location and scale constraints on this auxiliary distribution instead of on the original

mixture in the infinite case. But is it really that simple and if yes is it enough to just explain it this way?]

2.4.5 Related approaches in the literature

An alternative approach to dealing with the identifiability issues is to impose them after inference, for instance through appropriate transformations of the MCMC output produced with the unidentified model. An example of this is the treatment of the sign issue discussed earlier. This approach is often equivalent to assuming certain priors for the parameters of the factor model, which imply a nontrivial prior dependence among them. In some cases, the induced prior distribution can be derived analytically and exhibit desirable properties. For example in the framework of a factor model, Ghosh and Dunson (2009) show that this mechanism can be used to induce heavy-tailed priors on the factor loadings, which are well-defined and more flexible than the usual normal prior. In other cases, the implied prior dependence might be more difficult to grasp. This is for example the case in the paradigm of Yang et al. (2010), who work with the same transformation as ours, but rely on a post-processing stage to achieve identification. This posterior transformation implies a complicated prior on the loadings because of the transformation that involves a mixture of Gaussians. Instead, we impose the identifiability constraints *a priori*, but exploit the connection to the unidentifiable model to build efficient marginal data augmentation algorithms. We therefore use a different prior than Yang et al. (2010)'s. This difference is of major importance when it is crucial to know precisely the type of prior assumed on the loadings, for instance to implement stochastic search algorithms on the structure of the factor loading matrix, as already mentioned earlier (see section 2.2).

3 Marginal data augmentation methods for nonparametric factor models

3.1 Accelerating MCMC using unidentifiable model formulations

[OM: We need a bit more references here to connect to the literature I think][REMI: Added more references.]

Marginal Data Augmentation (MDA) methods (Meng and van Dyk, 1999) emerged in parallel with parameter-expansion methods (Liu and Wu, 1999), as a by-product of different attempts made to improve the convergence of the EM-algorithm (Meng and van Dyk, 1997; Liu et al., 1998). These approaches start from the observation that the introduction of extra parameters into the model (called *working parameters*), which cannot be identified from the data but can be sampled along the remaining parameters of the model, can dramatically improve convergence and mixing of the MCMC sampler. Based on this result, Meng and van Dyk (1999), van Dyk and Meng (2001), and van Dyk (2010) have formalized the mechanisms of MDA, and provided extensive examples to apply these methods to a wide range of models.

These approaches have proved to be particularly efficient in some types of models where convergence is usually very slow, to the point it can hinder proper inference, such as in latent variable models. For example, MDA methods have been successfully applied to a variety of discrete choice models, such as the multinomial probit (Imai and van Dyk, 2005; Xiyun and van Dyk, 2015), the multivariate probit (Lawrence et al., 2008), the multinomial logit (Scott, 2011), and also in factor analysis (Ghosh and Dunson, 2009; Yang et al., 2010; Frühwirth-Schnatter and Lopes, 2010; Conti et al., 2014), and to the sampling of correlation matrices (Liu and Daniels, 2006; Liu, 2008).

MDA methods provide the advantage of allowing to sample indirectly from complicated distributions that would otherwise be difficult to simulate. This feature is particularly useful in our framework: the Dirichlet process hierarchical model is challenging to simulate in its constrained version, but it can be marginally augmented to make it easier to handle. Last but not least, these methods are usually easy to implement—only a few additional working parameters need to be sampled at a low marginal cost, and no tuning is required. Hence, we can decouple the modelling, for which we can impose constraints for identifiability, from the computation, which can be done efficiently despite the complicated posteriors the modelling implies.

3.2 Working parameters for the nonparametric factor model

[REMI: All the transformed parameters and latent variables $\widetilde{\theta}$, $\widetilde{\Lambda}$, $\widetilde{\delta}$ are now also called 'working parameters'. I think only only $\widetilde{\mu}$ and $\widetilde{D}$ qualify as working parameters (only those two are integrated out, the other ones are just transformed).]

We build efficient MDA algorithms for the identifiable nonparametric factor model proposed in Section 2 using the following working parameters: $\widetilde{\theta}, \widetilde{\mu}, \widetilde{D}$, as they have already been defined in Section 2, and

$$\begin{aligned} \widetilde{\Lambda} &= \Lambda \widetilde{D}^{-\frac{1}{2}}, \\ \widetilde{\delta} &= \delta - \widetilde{\Lambda}\widetilde{\mu}. \end{aligned} \tag{17}$$

The backbone of the MDA algorithm we propose are the following results about the distribution of the working parameters. These are key to the efficient MCMC implementation we propose.

[REMI: Made minor modifications to proposition 1 (the beginning was still for the finite case only). I think it is now consistent and works for both finite and infinite cases.]

Proposition 1. *Consider the parameters $\widetilde{\mu}$ and $\widetilde{D}$ defined in eqs. (7) and (8), and the mappings from $\widetilde{\vartheta}$ to ϑ as defined in eqs. (9) and (10). Then, the normal-inverse-Wishart prior distribution specified on $\widetilde{\vartheta}_k = \{\widetilde{\mu}_k, \widetilde{\Phi}_k\}$ in eqs. (3) and (4), for $k \in \mathcal{K}$, implies that*

$$f(\widetilde{\mu}, \widetilde{D} \mid \vartheta, G, s_0, \nu_0, A_0) = f(\widetilde{\mu} \mid \widetilde{D}, \vartheta, G, A_0) \prod_{j=1}^{P} f(\widetilde{D}_j \mid \vartheta, G, \nu_0, s_0),$$

with

$$\widetilde{\mu} \mid \widetilde{D}, \vartheta, G, A_0 \sim \mathcal{N}_P\left(-\widetilde{D}^{\frac{1}{2}}\left(\sum_{k\in\mathcal{K}}\Phi_k^{-1}\right)^{-1}\left(\sum_{k\in\mathcal{K}}\Phi_k^{-1}\mu_k\right);\, A_0\widetilde{D}^{\frac{1}{2}}\left(\sum_{k\in\mathcal{K}}\Phi_k^{-1}\right)^{-1}\widetilde{D}^{\frac{1}{2}}\right), \tag{18}$$

$$\widetilde{D}_j \mid \vartheta, G, \nu_0, s_0 \sim \mathcal{IG}\left(\frac{\nu_0|\mathcal{K}|}{2};\, \frac{s_0}{2}\sum_{k\in\mathcal{K}}\left[\Phi_k^{-1}\right]_{[jj]}\right), \qquad \textit{for } j = 1,\ldots,P, \tag{19}$$

where $[\cdot]_{[jj]}$ *denotes the* j*th diagonal element of the corresponding matrix,* $\mathcal{K} = \{1,\ldots,K\}$ *in the finite mixture model,* $\mathcal{K} = \mathcal{I}^{(al)}$ *in the infinite mixture model and* $|\mathcal{K}|$ *is the cardinal number of the set* $\mathcal{K}$*.*

Proof. See appendix A.1. □

In the case of the finite mixture model, G can be dropped from the conditioning sets above. Interestingly, conditionally on ϑ, $(\widetilde{\mu}, \widetilde{D})$ are independent of the mixture probabilities p_k that are used in eqs. (7) and (8). However, in the infinite mixture model the construction imposes a prior dependence of $\widetilde{\mu}$ and $\widetilde{D}$ on G, although only via the active set $\mathcal{I}^{(\mathrm{al})}$ and the number of active components $|\mathcal{I}^{(\mathrm{al})}|$, implied by G.

The other distributions we will need for the implementation of the MDA algorithm are those that correspond to the parameters defined in eq. (17). However, it is a simple consequence of their definitions and the priors on the identifiable parameters in eqs. (13) and (14), that

$$f(\widetilde{\delta}, \widetilde{\lambda} \mid \widetilde{\mu}, \widetilde{D}, c_0, d_0) = f(\widetilde{\delta} \mid \widetilde{\lambda}, \widetilde{\mu}, \widetilde{D}, c_0)\prod_{q=1}^{Q} f(\widetilde{\lambda}_q \mid \widetilde{D}, d_0),$$

with:

$$\widetilde{\delta} \mid \widetilde{\lambda}, \widetilde{\mu}, \widetilde{D}, c_0 \sim \mathcal{N}\left(-\widetilde{\Lambda}\widetilde{\mu};\, c_0 I_Q\right), \tag{20}$$

$$\widetilde{\lambda}_q \mid \widetilde{D}, d_0 \sim \mathcal{N}\left(0;\, d_0\widetilde{D}^{-1}\right), \tag{21}$$

for $q = 1,\ldots,Q$.

3.3 MDA sampling scheme

For the MDA sampling scheme to produce a posterior sample of the parameters of the identified model, it is important to sample the working parameters jointly with the parameters of interest (van Dyk, 2010). The sampler is presented as algorithm 1 below. Only the parameters and latent variables with an exponent (t) are kept for posterior inference. The other ones are auxiliary draws that are immediately discarded at the end of the corresponding MCMC iteration. Some of them, like the working parameters $\widetilde{\mu}$ and $\widetilde{D}$, may

be updated several times in a single MCMC iteration. In that case, their most up-to-date values are used in any given substep of the MCMC sampler.

[OM: Remember to add alpha in all this][REMI: done.]

[REMI: In the previous version I made explicit the sampling of the working parameters in steps 2) and 3). You removed it. Both formulations work, the new one makes it clearer that the working parameters are integrated out, and notation is lighter. I only removed the tildes on δ and Λ in step 3) to be consistent with the rest.]

[REMI: I moved the computation of $\widetilde{\mu}$ and $\widetilde{D}$ from step 4 to step 8. This does not change the sampler (as the working parameters are not used in steps 5-7) and I think it looks more natural this way (steps 4-6 are all performed in the unidentified model, and in the infinite case it was weird to do this computation in step 4, since additional mixture components can be introduced retrospectively in step 5-6, thus changing the resulting working parameters).]

Algorithm 1. MDA sampler

Initialization. Assign starting values for all parameters $\{\delta^{(0)}, \Lambda^{(0)}, \Sigma^{(0)}, p^{(0)}, \vartheta^{(0)}\}$ and latent variables $\{\theta^{(0)}, G^{(0)}\}$.

MCMC sampling. At each iteration $t = 1, \ldots, T$, cycle through the following steps:

1) Sample $\Sigma^{(t)}$ from $f(\Sigma \mid Y, \delta^{(t-1)}, \Lambda^{(t-1)}, \theta^{(t-1)})$.

2) Sample $\widetilde{\theta}$ from $f(\widetilde{\theta} \mid Y, \delta^{(t-1)}, \Lambda^{(t-1)}, \Sigma^{(t)}, G^{(t-1)}, \vartheta^{(t-1)})$, in steps:

- **a)** Sample $\widetilde{\mu}$ and $\widetilde{D}$ from $f(\widetilde{\mu}, \widetilde{D} \mid \vartheta^{(t-1)})$, see eqs. (18) and (19).
- **b)** Sample θ from $f(\theta \mid Y, \delta^{(t-1)}, \Lambda^{(t-1)}, \Sigma^{(t)}, G^{(t-1)}, \vartheta^{(t-1)})$, then compute $\widetilde{\theta}_i = \widetilde{\mu} + \widetilde{D}^{\frac{1}{2}}\theta_i$, for $i = 1, \ldots, N$.

3) Sample $\delta^{(t)}$, $\Lambda^{(t)}$ from $f(\delta, \Lambda \mid Y, \widetilde{\theta}, \Sigma^{(t)}, G^{(t-1)}, \vartheta^{(t-1)})$ in steps:

- **a)** Sample $\widetilde{\mu}, \widetilde{D}$ from $f(\widetilde{\mu}, \widetilde{D} \mid \widetilde{\theta}, G^{(t-1)}, \vartheta^{(t-1)})$.
- **b)** Sample $\widetilde{\Lambda}$ from $f(\widetilde{\Lambda} \mid Y, \widetilde{\theta}, \Sigma^{(t)}, \widetilde{\mu}, \widetilde{D})$.
- **c)** Sample $\widetilde{\delta}$ from $f(\widetilde{\delta} \mid Y, \widetilde{\theta}, \Sigma^{(t)}, \widetilde{\Lambda}, \widetilde{\mu}, \widetilde{D})$.
- **d)** Compute and save $\Lambda^{(t)} = \widetilde{\Lambda}\widetilde{D}^{\frac{1}{2}}$ and $\delta^{(t)} = \widetilde{\delta} + \widetilde{\Lambda}\widetilde{\mu}$.

4) Sample $\widetilde{\vartheta}$ from $f(\widetilde{\vartheta} \mid \widetilde{\theta}, G^{(t-1)})$.

5) Sample $G^{(t)}$ from $f(G \mid \widetilde{\theta}, p^{(t-1)}, \widetilde{\vartheta})$.

6) Sample $V_k^{(t)}$ from $f(V_k \mid G^{(t)}, \alpha^{(t-1)})$, and compute the corresponding mixture weights p_k.

7) Sample $\alpha^{(t)}$ from $f(\alpha \mid G^{(t)})$.

8) Obtain $\widetilde{\mu}$ and $\widetilde{D}$ from $\widetilde{\vartheta}$ sampled in step 4, using eqs. (7) and (8). Apply the transformation in eqs. (9) and (10) to produce the parameters $\vartheta^{(t)}$ corresponding to the identified model. Transform the latent factors back to the identified model as $\theta_i^{(t)} = \widetilde{D}^{-\frac{1}{2}}\left(\widetilde{\theta}_i - \widetilde{\mu}\right)$, for $i = 1, \ldots, N$.

Post-processing. Perform a sign switch on the factor loading matrix, mixture means and mixture covariances, to ensure that the model is identified with respect to the signs of the latent factors and factor loadings.[a]

[a] See Frühwirth-Schnatter and Lopes (2010) and Conti et al. (2014). Signs are switched such that the first nonzero elements in each column of Λ are always positive across MCMC iterations.

The main difference between this MDA sampling scheme and a standard Gibbs sampler for factor models is the additional working parameters that need to be sampled along with

the parameters of interest and the latent variables of the model. These additional steps, however, only represent a marginal additional cost. The intermediate values of the working parameters are all drawn directly from standard distributions, except at step 3-a, where a Metropolis-Hastings step is implemented. For details on the sampler, see appendix B.

Since no information on the working parameters can be retrieved from the data, they are sampled from their conditional prior distribution the first time they are required, in step 2a. The latent factors θ are then sampled from the identified model, and immediately transformed to obtain their counterpart in the expanded model. This step is equivalent to sampling directly from $f(\widetilde{\theta} \mid Y, \delta, \Lambda, \Sigma, G, \vartheta)$.

The finite and infinite cases proceed differently in steps 4 to 6. All these substeps are performed in the unidentified model, but the mixture parameters $\widetilde{\vartheta}_k$ are updated slightly differently. In the finite case, the mixture parameters are sampled from their posterior distribution for the alive mixture components and from their prior for the empty components. In the infinite case, this is not feasible. Instead, only the mixture parameters of the alive components are sampled in step 4, and any new components that may be required in the subsequent steps 5 and 6 to increase the size of the mixture are introduced *retrospectively*, using the procedure of Papaspiliopoulos and Roberts (2008). Since the mixture parameters are all updated in the unidentified model, the prior dependence on G that affects ϑ in the infinite case is not relevant at these stages. This dependence will be later restored by the transformation in step 8. Therefore, these steps represent a standard Gibbs step in the finite case, and a standard—but non-trivial—retrospective sampling step in the infinite case.[7]

The parameter transformation carried out in step 8 guarantees that the mixture parameters fulfill the identification requirements *exactly* at each step of the MCMC sampler. Importantly, the parameters and latent variables that are affected by the expansion are always sampled simultaneously with the working parameters. This ensures that the sampling scheme preserves the prior distribution of the parameters in the identified model, and does not distort the posterior distribution, as would happen if the sampling was done conditional on the working parameters.

[OM: This needs additions, see email] [REMI: Does this work better now?]

4 Illustrations with synthetic and real data

To investigate the performance of our approach in practice, we run our sampler on simulated and real data. We compare the results obtained with different algorithms for the infinite Dirichlet process.

[7] See details in appendix B.6.

4.1 Simulation Study

4.1.1 Model specification

A synthetic data set with $Q = 9$ observed measurements and $N = 2,000$ observations is generated from the factor model specified in eq. (1), for $P = 2$ latent factors and using the following parameter values for the structural part of the model:

$$\begin{aligned} \delta' &= \begin{pmatrix} -1.2 & -0.9 & -0.6 & -0.3 & 0.0 & 0.3 & 0.6 & 0.9 & 1.2 \end{pmatrix}, \\ \Lambda' &= \begin{pmatrix} 1.0 & 0.9 & 0.8 & 0.0 & 0.0 & 0.0 & 0.8 & 0.6 & 0.4 \\ 0.0 & 0.0 & 0.0 & 1.0 & 0.9 & 0.8 & 0.4 & 0.6 & 0.8 \end{pmatrix}, \\ \Sigma &= \operatorname{diag}\begin{pmatrix} 0.05 & 0.20 & 0.40 & 0.05 & 0.20 & 0.40 & 0.05 & 0.20 & 0.40 \end{pmatrix}. \end{aligned} \tag{22}$$

Each factor has three dedicated measurements, and the last three measurements load on both factors. This type of structure is very common in the social sciences, where some particular tests are designed to measure specific traits (think of an IQ test), while others capture several features simultaneously (e.g., personality tests measuring self-esteem and self-confidence at the same time). The idiosyncratic variances in Σ are unbalanced to vary the proportion of noise affecting each measurement.

The distribution of the latent factors is specified as a mixture of three normal distributions, using the following parameter values in the expanded version of the model:

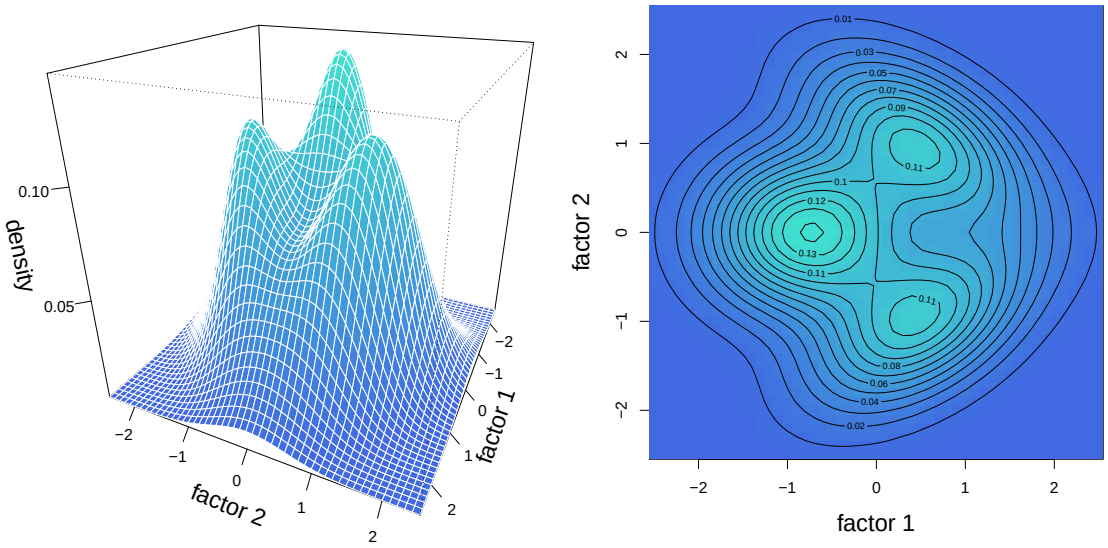
$$\begin{aligned} p_1 &= 0.4, & p_2 &= 0.3, & p_3 &= 0.3, \\ \widetilde{\mu}_1 &= \begin{pmatrix} 0 & 0 \end{pmatrix}, & \widetilde{\mu}_2 &= \begin{pmatrix} 1.4 & -1.4 \end{pmatrix}, & \widetilde{\mu}_3 &= \begin{pmatrix} 1.4 & 1.4 \end{pmatrix}, \\ \widetilde{\Phi}_1 &= \begin{pmatrix} 0.7 & 0.0 \\ 0.0 & 0.7 \end{pmatrix}, & \widetilde{\Phi}_2 &= \begin{pmatrix} 0.8 & 0.4 \\ 0.4 & 0.8 \end{pmatrix}, & \widetilde{\Phi}_3 &= \begin{pmatrix} 0.8 & -0.4 \\ -0.4 & 0.8 \end{pmatrix}. \end{aligned}$$

The parameters of the mixture are transformed according to **????????** to standardize the latent factors to be centered around zero and have unit variances. The joint distribution is displayed in fig. 1 and exhibits three modes, corresponding to the three components of the mixture. Despite its shape that may look exotic at first sight, it may not be unlikely to encounter such a distribution in practice, where for a low level of the first trait θ_1, the population has a unimodal conditional distribution on the other trait $(\theta_2 \mid \theta_1)$, while this conditional distribution becomes bimodal on the other end of the distribution of the first trait θ_1. Common methods traditionally used in the empirical literature (i.e., standard factor analysis) do not allow to deal with such distributions, thus potentially leading to biased results.

4.1.2 Identification and prior specification

Following the discussion in section 2.2, each latent factor needs at least two dedicated measurements for nonparametric identification. The true factor loading matrix Λ specified

Figure 1: True joint distribution of the latent factors in the simulation study.



in eq. (22) has three measurements loading exclusively on each factor, which is sufficient to identify the model. We constrain the corresponding elements to zero for this purpose.

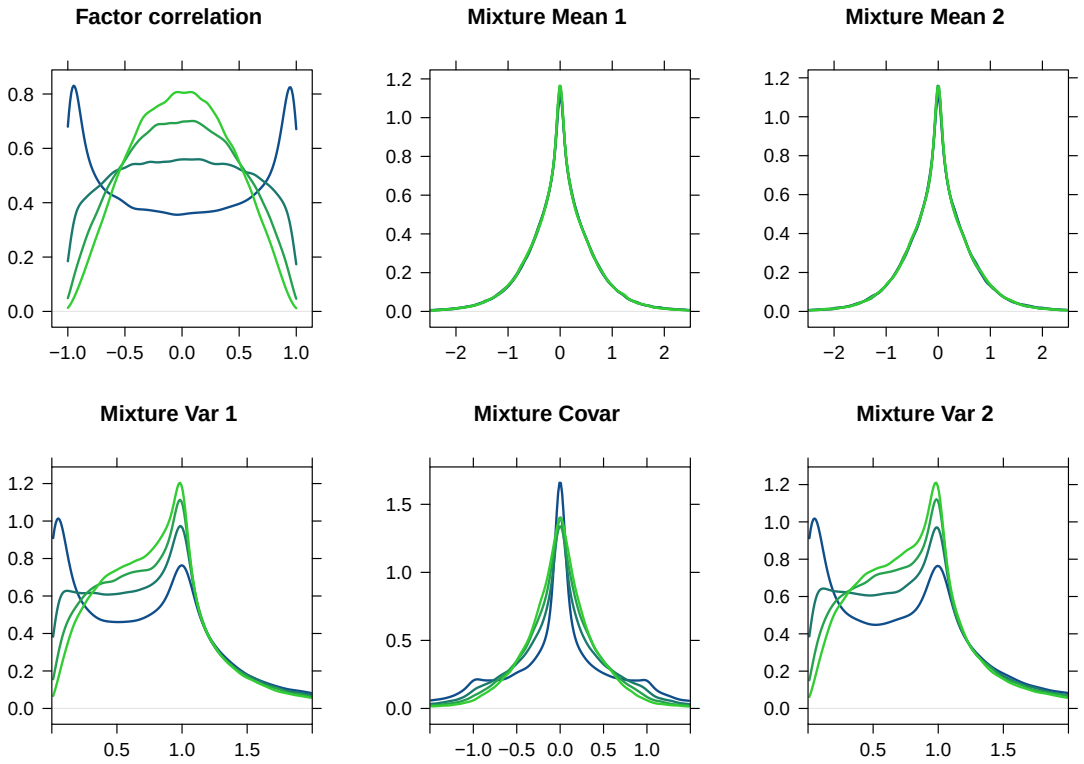
The prior distribution of the mixture parameters has a complicated form in the identified model—see eq. (A1). It can, however, easily be simulated to understand the role of these prior parameters. The two most important prior parameters are the degrees of freedom ν_0 of the mixture covariance matrices and the scale A_0 of the prior covariance matrix of the mixture means. Figures 2 and 3 display the prior distributions of the mixture parameters in the identified model, as well as of the corresponding correlation between the latent factors, in a model with 2 factors and a mixture with 2 components.

In our simulation study, we specify the prior parameters as follows:

$$\begin{aligned} c_0 &= 10.0, & d_0 &= 10.0, & a_0 &= 2.0, & b_0 &= 1.0, \\ A_0 &= 1.0, & \nu_0 &= 3, & s_0 &= 1.0, & \alpha &\in \{0.2, 0.6, 1.0, 1.2\}. \end{aligned}$$

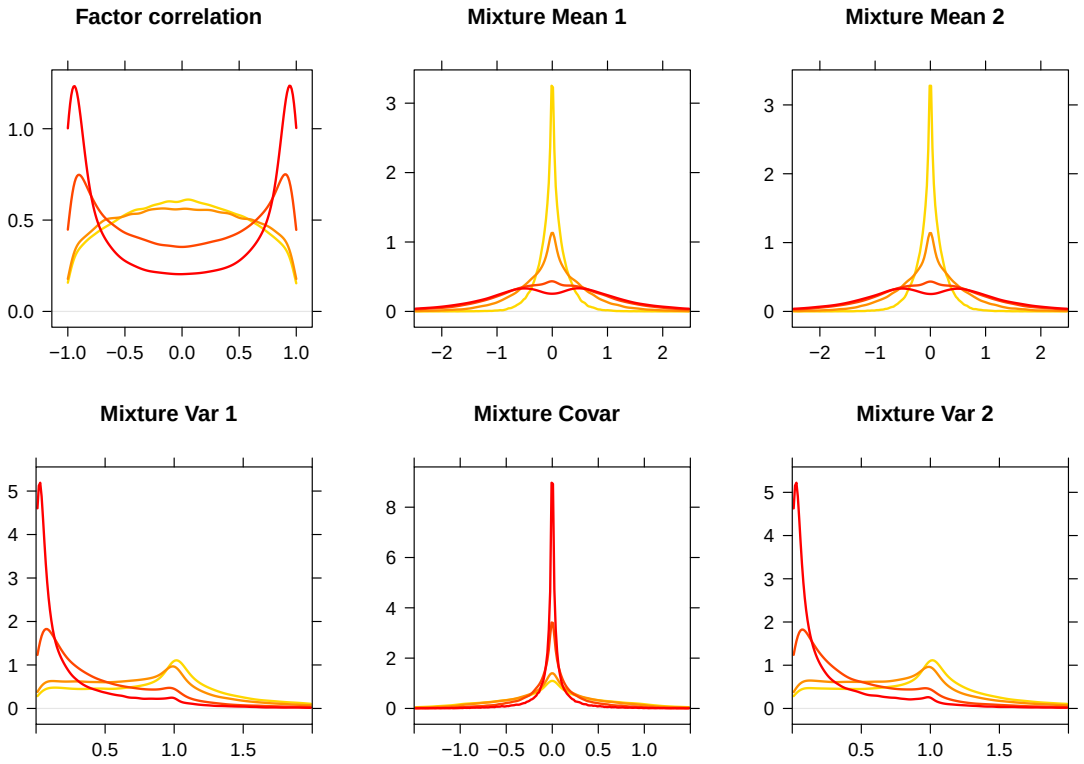
The distribution on the mixture parameters does not impose any strong prior—setting $\nu_0 = P + 1$ for the degrees of freedom of the covariance matrices implies a uniform distribution on the prior correlation between the latent factors. For the concentration parameter of the Dirichlet process α, we try different values to see the impact on the number of mixture components. Figure 4 (top row) shows how this parameter influences *a priori* the number of mixture components. The larger α, the more likely new mixture components will be introduced into the process *a priori*. This prior parameter can therefore be tuned to control the expansion of the Dirichlet process in terms of number of mixture components. This is analogous to alternative nonparametric approaches, such as kernel density estimation methods, where a smoothing parameter usually needs to be selected by the analyst to control the level of smoothness of the estimator (e.g., bandwidth).

Figure 2: Implied prior distribution on mixture parameters in identified model, for different degrees of freedom $\nu_0 = 2$, 3, 4 and 5 (from dark blue to green), $A_0 = 1$, $s_0 = 1$, and $\tau_0 = 1$.



Note: The mixture is symmetric (label-switching), therefore only the parameters of one mixture component are displayed here. Densities obtained from 100,000 draws random.

Figure 3: Implied prior distribution on mixture parameters in identified model, for different scale parameters $A_0 = 0.1,\ 1,\ 10,\ 50$ (from gold to red), $\nu_0 = 3$, $s_0 = 1$, and $\tau_0 = 1$.



Note: The mixture is symmetric (label-switching), therefore only the parameters of one mixture component are displayed here. Densities obtained from 100,000 draws random.

4.1.3 MCMC tuning

The sampler is run for 55,000 iterations, where the first 5,000 iterations are discarded as burn-in period. A sign-switching is performed *a posteriori* on the factor loading matrix, mixture means and mixture covariances, to ensure that the model is identified with respect to the signs of the latent factors and factor loadings (see Frühwirth-Schnatter and Lopes, 2010; Conti et al., 2014).[8] This simple transformation is innocuous for the interpretation of the results. [OM: remove this from here, it is elsewhere]

4.1.4 Simulation results

We first present and discuss how our algorithm manages to retrieve the distribution of the latent factors, before taking a closer look at the statistical properties of the sampler.

The number of mixture parameters is not specified *a priori* but sampled along with the other parameters of the model. The concentration parameter α of the Dirichlet process controls the level of smoothing of the nonparametric approach. The larger α, the larger the number of mixture components. This result holds both *a priori* and *a posteriori*: fig. 4 shows that with $\alpha = 0.2$, the number of components is a posteriori equal to the true one, while with $\alpha = 1.2$ too many components are introduced into the Dirichlet process, resulting in overfitting. Most of these extra components, however, appear to be just noise, as revealed by the third row of this figure where only mixture components with at least 5 observations are considered. Figure 5 shows that the sampler provides a very good mixing with respect to the number of mixture components. The remaining results are discussed for the case $\alpha = 0.2$.

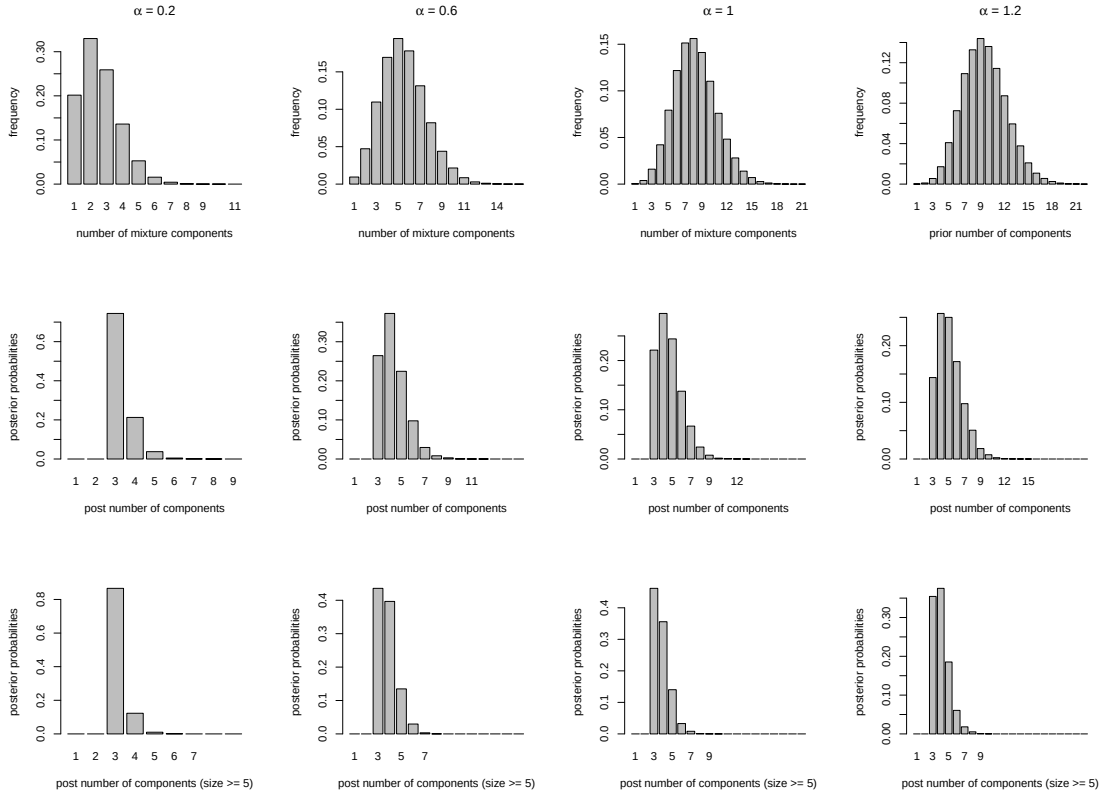
The fit of the estimated distribution to the true distribution is displayed in fig. 6. The top panel of this figure shows that the algorithm manages to recover the true distribution of the latent factors, compared to fig. 1. The four figures at the bottom of fig. 6 compare the posterior distribution obtained from the algorithm to the true distribution for different slices of the joint distribution (i.e., conditional distributions $f(\theta_1 \mid \theta_2)$ and $f(\theta_2 \mid \theta_1)$), together with the corresponding 90% confidence intervals. These figures show an excellent fit for this type of mixture with a data set of 2,000 observations.

As for the structural parameters of the factor model, the results also look very encouraging. Figure 7 shows the trace plots, posterior distributions and autocorrelograms of some selected parameters—intercept term, two factor loadings and idiosyncratic of the last equation, respectively. The sampler exhibits a very good mixing, as indicated by the trace plots and the autocorrelations that converge to zero very quickly. These results are worth pointing out, as latent variable models are often plagued by slow convergence and bad mixing.[9]

[8] Signs are switched such that the first nonzero elements in each column of Λ are always positive across MCMC iterations.

[9] To remedy this problem, analysts usually increase the number of MCMC iterations and resort to a thinning of the Markov chain, which is not necessary here given the good mixing.

Figure 4: Prior, posterior and truncated (only components with at least 5 observations) posterior numbers of mixture components for $\alpha = 0.2$, 0.6, 1.0 and 1.2, for the retrospective sampler.



Note: Simulations based on 100,000 replications for the prior distribution.

Figure 5: Trace plot of number of mixture components, retrospective sampler, simulations with $\alpha = 0.2$, 0.6, 1.0 and 1.2 (from top to bottom).

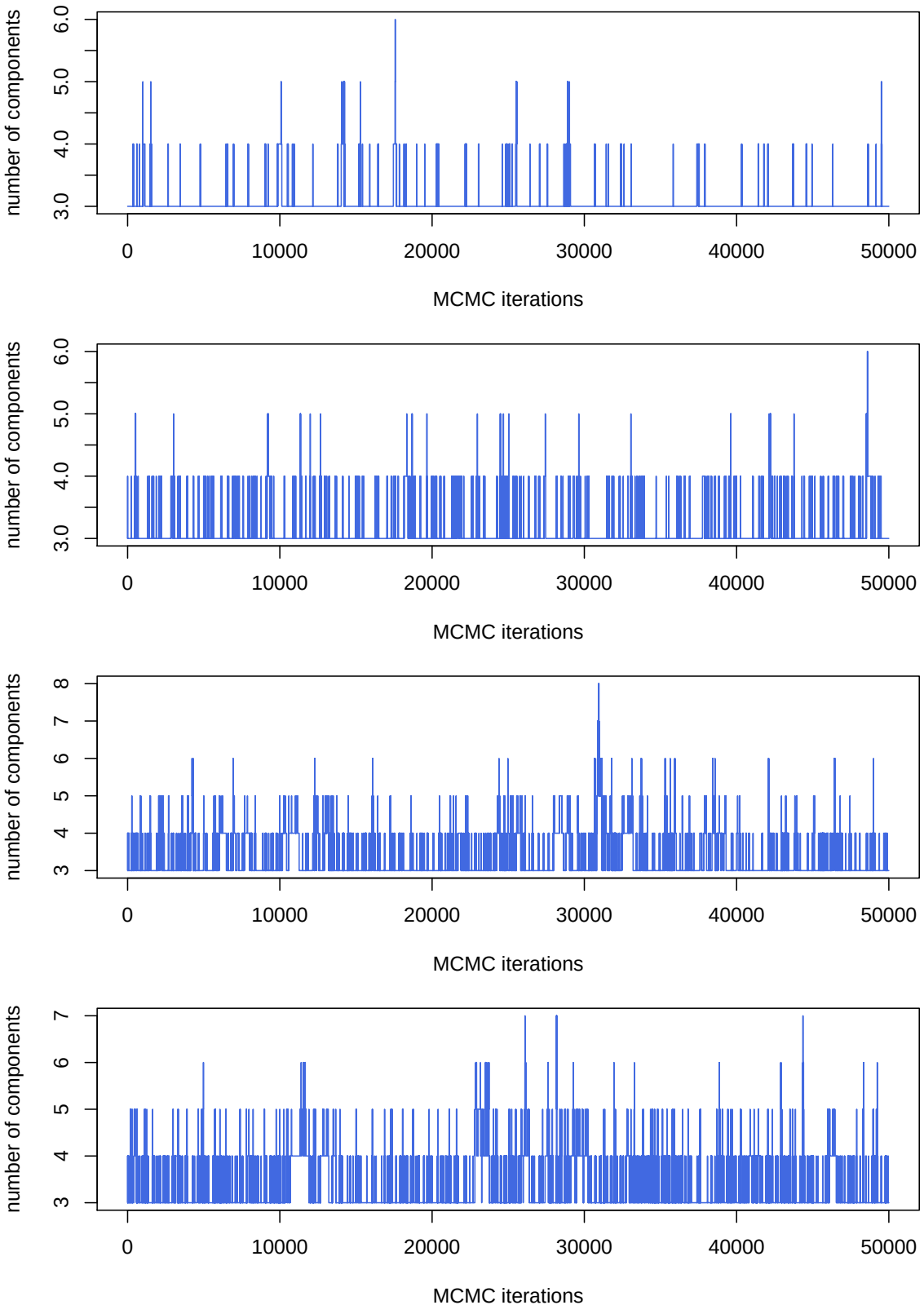


Figure 6: Restrospective MCMC with $\alpha = 0.2$ — Posterior joint distribution of the latent factors. 90% highest posterior density interval in dashed line. True distribution in blue.

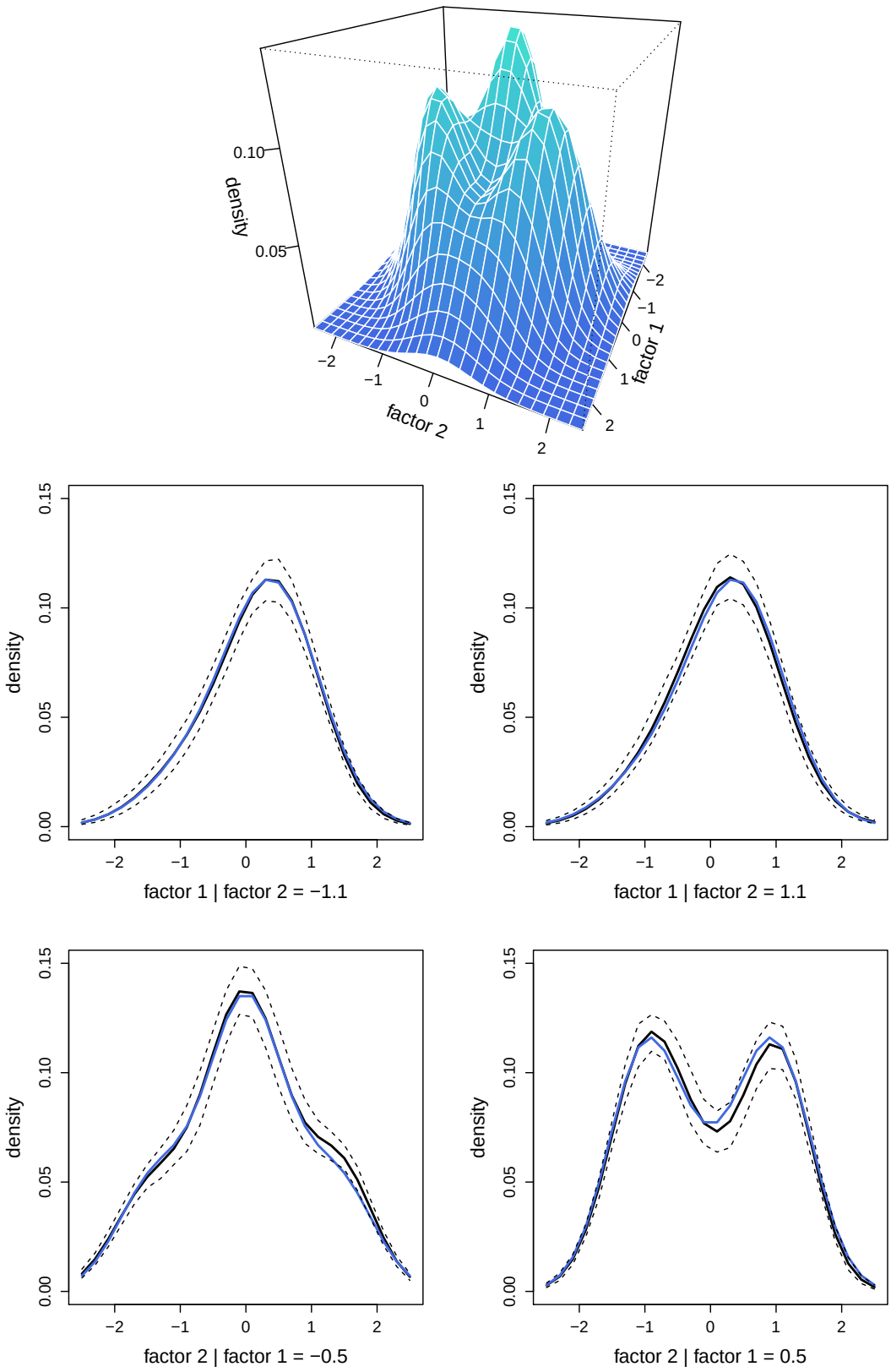
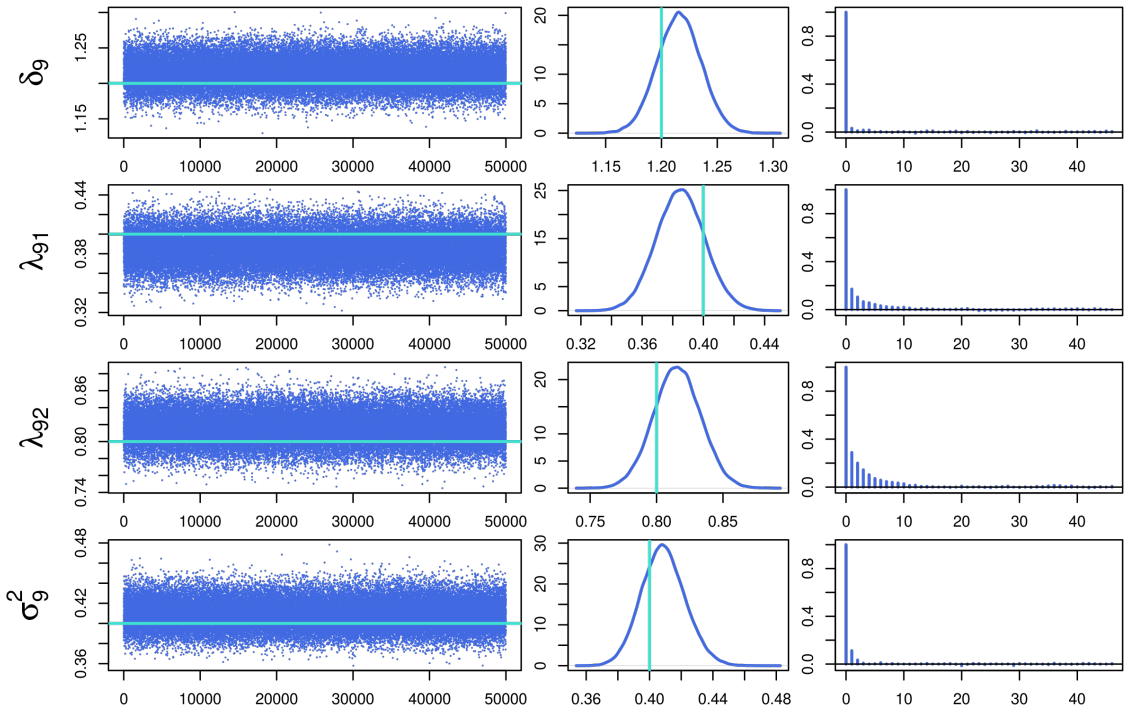


Figure 7: Posterior results for selected parameters. Trace plots, posterior densities and autocorrelograms. True values in turquoise.



This good behavior of the sampler is obtained thanks to the MDA procedure that boosts MCMC. The sampler does not produce independent and identically distributed draws though, as can be seen from the autocorrelations that need a few iterations to disappear. This can be explained by the fact the prior distribution of the working parameters introduced to expand the model is a *conditional* prior distribution, which includes the mixture parameters of the identified model ϑ in its conditioning set, see eqs. (18) and (19). This does not invalidate the approach (van Dyk, 2010), but slightly deteriorates the efficiency of the algorithm, because of the dependence introduced among the parameters across MCMC iterations—the working parameters are always sampled conditional on the value of ϑ from the previous MCMC iteration.

Finally, we take a closer look at the efficiency of our approach, by comparing different algorithms for the sampling of the Dirichlet process. Besides the retrospective MCMC sampler (`retro`), we implement a constrained version of the Dirichlet process (`trunc`, where the Dirichlet process is truncated to a maximum of 10 components, see Ishwaran and James, 2002), as well as the last two algorithms for the unconditional Dirichlet process proposed by Neal (2000), `neal7` and `neal8`, where the latter is run with $m = 3$ intermediate states. We perform this comparison for several values of the concentration parameter α, and compare the inefficiency factors for different statistics of interest: the number of active mixture components K (i.e., non-empty components), some selected parameters (intercept term, nonzero factor loading and idiosyncratic variance corresponding to the first measurement), as well the deviance of the estimated distribution of the latent factors

θ and of the observed measurements Y, calculated as:

$$D(\theta) = -2\sum_{i=1}^{N} \log \left\{ \sum_{k \in \mathcal{I}^{(al)}} \frac{N_k}{N} f(\theta_i \mid \mu_k, \Phi_k) \right\},$$

$$D(Y) = -2\sum_{i=1}^{N} \log \left\{ \sum_{k \in \mathcal{I}^{(al)}} \frac{N_k}{N} f(Y_i \mid \delta, \Lambda, \Sigma, \mu_k, \Phi_k) \right\}.$$

The deviance, a function of several relevant model parameters, has been used as a measure of fit by Neal (2000), Green and Richardson (2001), and Papaspiliopoulos and Roberts (2008) in their comparison studies. The inefficiency factor is a popular statistic used to monitor the convergence and the mixing of the Markov chain. Computed as the inverse of the relative numerical efficiency (Geweke, 1989), it provides a measure of the number of MCMC iterations that is required by the sampler to provide the same numerical precision as an hypothetical i.i.d. sample from the target distribution.[10]

The results of this comparative study are presented in table 1 and show that all the approaches provide similar results in terms of numerical efficiency. The inefficiency factors are very low for this type of model—another confirmation of the benefit of the MDA approach. The sampler is slightly less efficient for the factor loadings (see column for λ_{11}), due to dependence of the working prior on the identified parameters of the mixture, which introduces some persistence across MCMC iterations. This translates into a minor deterioration of the inefficiency factor for the deviance of the observed measurements, $D(Y)$, compared to the deviance of the latent factors $D(\theta)$, as the former is a function of the factor loadings. Last, increasing the concentration parameter α results in overfitting, which has a negative impact on numerical efficiency.

Overall, these simulation results look very promising. The sampler manages to estimate very precisely the distribution of the latent factors, with no particular prior information on its shape, kurtosis or skewness, and the structural parameters of the factor models are estimated very efficiently.

4.2 Empirical example

Many empirical applications in economics assume normality of the latent factors. With this assumption, inference is usually straightforward to carry out, and interpretation is facilitated. It is, however, legitimate to question the relevance of this assumption in practice. Does it make a difference for the inference of the latent part of the model to relax normality? To illustrate this problem, we partially revisit in this section the real data example presented in Conti et al. (2014), in which the latent factors are assumed to be Gaussian. How realistic is this assumption? Is it supported by the data?

[10] Lower inefficiency factors are better. For example, with an inefficiency factor of 5, 50,000 draws are required to provide a numerical precision equivalent to the one that could ideally be obtained with 10,000 i.i.d. draws.

Table 1: Inefficiency factors for the number of mixture components K, the deviance D of estimated distributions of the latent factors θ and of the observed measurements Y, and for some selected parameters.

	K	$D(\theta)$	$D(Y)$	δ_1	λ_{11}	σ_1^2
			$\alpha = 0.2$			
`retro`	31.56	1.12	2.68	1.06	6.70	2.52
`trunc`	30.22	1.04	2.85	1.04	6.58	2.61
`neal7`	18.96	1.07	2.89	1.04	6.54	2.53
`neal8`	14.58	2.39	2.85	1.04	6.91	2.43
			$\alpha = 1.0$			
`retro`	39.71	2.00	3.58	1.14	7.24	3.08
`trunc`	40.50	2.59	3.44	1.18	7.16	2.42
`neal7`	25.79	1.74	3.43	1.17	7.18	2.93
`neal8`	15.36	1.95	3.38	1.13	7.31	2.71
			$\alpha = 2.0$			
`retro`	43.00	5.44	3.71	1.27	8.16	2.92
`trunc`	34.25	4.90	4.61	1.23	7.46	2.93
`neal7`	23.27	1.53	4.20	1.31	8.13	2.92
`neal8`	21.30	2.70	4.03	1.27	7.74	3.06

4.2.1 Data

The data set is drawn from the British Cohort Study (BCS), a longitudinal survey that follows all babies born in one particular week of April 1970 in the United Kingdom. It includes a large number of measurements on cognitive abilities, socio-emotional traits, behavioral and physical development at different stages in the life cycle of the surveyed individuals, and therefore represents a unique opportunity for psychologists and economists to study human capital development.

While the original analysis in Conti et al. (2014) uses a set of 131 measurements and extracts 13 latent factors with a stochastic search on the latent structure of the model, in this paper we restrict our analysis to two dimensions and focus on cognitive ability and behavioral problems. The first one is measured by 7 test scores, while the second is captured by 16 measurements related to the Rutter and Conners scales.[11]

4.2.2 Inference

We do not incorporate any strong prior information into the model and use the same prior specification as in the simulation study. The Dirichlet process is tailored to favor parsimonious solutions with smaller numbers of mixture components, with a concentration parameter α set to 0.2. The sensitivity of the results to this tuning parameter is investigated by increasing its value up to 1.2, as in the simulation study. As expected,

[11] The interested reader is referred to the original paper (Conti et al., 2014) for full details.

larger values result in solutions with more mixture components, but without changing the shape of the estimated distribution, nor the overall fit of the model.

To identify the structural part of the model, a dedicated structure is assumed, where each measurement loads on its corresponding latent factor. We therefore create two clusters of measurements, where the underlying factors are allowed to be correlated. The sampler is run for 55,000 iterations, where only the last 50,000 ones are kept for posterior inference.

?? displays the posterior joint distribution of the latent factors. This distribution exhibits several modes, and fat tails for larger values of the behavioral problems factor. The multiplicity of the number of modes is confirmed by **??**, which shows that the algorithm visits models with 4 mixture components the most often, with a posterior probability larger than 0.7. Overall, the retrospective sampler switches very quickly between models with different numbers of mixture components, as shown in **??**.

This simple example reveals that the normality assumption is likely to be violated in this data set. Relaxing it allows the sampler to explore alternative solutions with non-standard distributions, which are more likely to be supported by the data. This misspecification of the model is likely to contaminate the interpretation of the results, and to affect the estimation if this model is subsequently used to measure the impact of these latent factors on economic outcomes.

5 Conclusion

This paper introduces a new approach to factor analysis with non-normal factors that draws on the literature on Bayesian nonparametric methods. It extends these approaches by placing the formal identification of the factor model at the core of the inferential procedure, guaranteeing that the algorithm only produces identified models during sampling. This is achieved by implementing a new sampling scheme for mixtures of normals based on marginal data augmentation, combined with a retrospective MCMC sampler for the Dirichlet process mixture model.

A simulation study is carried out and provides results that are very encouraging. The sampler is successful in retrieving the distribution of the latent factors nonparametrically, and exhibits very good properties for the structural part of the model in terms of convergence and mixing. A real data example illustrates the relevance of the method, by challenging the specification of an empirical study that implements a factor model with Gaussian factors. It appears that the normality assumption can be seriously questioned in practice. It remains an open question to investigate how a possible misspecification of the distribution of the latent factors can affect the inference of the model, and the conclusions of the practitioners. These important questions will be further investigated in future projects to obtain a full picture of the problem.

References

Aguilar, O., and M. West. 2000. "Bayesian Dynamic Factor Models and Portfolio Allocation". *Journal of Business & Economic Statistics* 18 (3): 338–357. doi:10.3905/jpm.1985.409020.

Almlund, M., A. L. Duckworth, J. J. Heckman, and T. Kautz. 2011. "Personality Psychology and Economics". Chap. 1 in *Handbook of the Economics of Education*, ed. by E. A. Hanushek, S. Machin, and L. Woessmann, 4:1–181. 2008. North-Holland, Elsevier. doi:10.1016/B978-0-444-53444-6.00001-8.

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*, ed. by J. Neyman, 5:111–150. Berkeley: University of California Press.

Antoniak, C. E. 1974. "Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems". *The Annals of Statistics* 2 (6): 1152–1174.

Attias, H. 1999. "Independent Factor Analysis". *Neural Computation* 11 (4): 803–51.

Bernanke, B. S., J. Boivin, and P. Eliasz. 2005. "Measuring the Effects of Monetary Policy: A Factor-Augmented Vector Autoregressive (Favar) Approach". *The Quarterly Journal of Economics* 120 (1): 387–422. doi:10.1162/0033553053327452.

Bhattacharya, A., and D. B. Dunson. 2011. "Sparse Bayesian Infinite Factor Models". *Biometrika* 98 (2): 291–306. doi:10.1093/biomet/asr013.

Carneiro, P., K. T. Hansen, and J. J. Heckman. 2003. "Estimating Distributions of Treatment Effects with an Application to the Returns to Schooling and Measurement of the Effects of Uncertainty on College Choice". *International Economic Review* 44 (2): 361–422. doi:10.1111/1468-2354.t01-1-00074.

Carvalho, C. M., J. Chang, J. E. Lucas, J. R. Nevins, Q. Wang, and M. West. 2008. "High-Dimensional Sparse Factor Modeling: Applications in Gene Expression Genomics". *Journal of the American Statistical Association* 103 (484): 1438–1456.

Conti, G., S. Frühwirth-Schnatter, J. J. Heckman, and R. Piatek. 2014. "Bayesian Exploratory Factor Analysis". *Journal of Econometrics* 183 (1): 31–57. doi:10.1016/j.jeconom.2014.06.008.

Cunha, F., and J. J. Heckman. 2008. "Formulating , Identifying and Estimating the Technology of Cognitive and Noncognitive Skill Formation". *Journal of Human Resources* 43 (4): 738–782.

Cunha, F., J. J. Heckman, and S. M. Schennach. 2010. "Estimating the Technology of Cognitive and Noncognitive Skill Formation." *Econometrica* 78 (3): 883–931. doi:10.3982/ECTA6551.

Escobar, M. D., and M. West. 1995. "Bayesian Density Estimation and Inference Using Mixtures". *Journal of the American Statistical Association* 90 (430): 577–588.

Fokoué, E., and D. M. Titterington. 2003. "Mixtures of Factor Analysers. Bayesian Estimation and Inference by Stochastic Simulation". *Machine Learning* 50:73–94.

Forni, M., and L. Gambetti. 2010. "The dynamic effects of monetary policy: A structural factor model approach". *Journal of Monetary Economics* 57 (2): 203–216. doi:10.1016/j.jmoneco.2009.11.009.

Frühwirth-Schnatter, S., and H. F. Lopes. 2010. "Parsimonious Bayesian Factor Analysis when the Number of Factors is Unknown". *Working Paper*:The University of Chicago Booth School of Business.

Geweke, J. F. 1989. "Bayesian Inference in Econometric Models Using Monte Carlo Integration". *Econometrica* 57 (6): 1317–1339.

Geweke, J. F., and G. Zhou. 1996. "Measuring the Pricing Error of the Arbitrage Pricing Theory". *Review of Financial Studies* 9 (2): 557–587.

Ghosh, J., and D. B. Dunson. 2009. "Default Prior Distributions and Efficient Posterior Computation in Bayesian Factor Analysis". *Journal Of Computational And Graphical Statistics* 18 (2): 306–320. doi:10.1198/jcgs.2009.07145.

Green, P. J., and S. Richardson. 2001. "Modelling Heterogeneity With and Without the Dirichlet Process". *Scandinavian Journal of Statistics* 28 (1999): 355–375.

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:doi:10.1016/j.jeconom.2003.10.011.

Heckman, J. J., J. Stixrud, and S. Urzúa. 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. 2005. "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.

Ishwaran, H., and L. F. James. 2001. "Gibbs Sampling Methods for Stick-Breaking Priors". *Journal of the American Statistical Association* 96 (453): 161–173.

— . 2002. "Approximate Dirichlet Process Computing in Finite Normal Mixtures". *Journal of Computational and Graphical Statistics* 11 (3): 508–532. doi:10.1198/106186002411.

Koopmans, T. C., and O. Reiersøl. 1950. "The Identification of Structural Characteristics". *The Annals of Mathematical Statistics* 21 (2): 165–181.

Lawrence, E., D. Bingham, C. Liu, and V. N. Nair. 2008. "Bayesian Inference for Multivariate Ordinal Data Using Parameter Expansion". *Technometrics* 50 (2): 182–191. doi:10.1198/004017008000000064.

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

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

Liu, X. 2008. "Parameter Expansion for Sampling a Correlation Matrix: An Efficient GPX-RPMH Algorithm". *Journal of Statistical Computation and Simulation* 78 (11): 1065–1076. doi:10.1080/00949650701519635.

Liu, X., and M. J. Daniels. 2006. "A New Algorithm for Simulating a Correlation Matrix Based on Parameter Expansion and Reparameterization". *Journal of Computational and Graphical Statistics* 15 (4): 897–914. doi:10.1198/106186006X160681.

Lopes, H. F., and M. West. 2004. "Bayesian Model Assessment in Factor Analysis". *Statistica Sinica* 14:41–67.

Lucas, J. E., C. M. Carvalho, Q. Wang, A. Bild, J. Nevins, and M. West. 2006. "Sparse Statistical Modelling in Gene Expression Genomics". In *Bayesian Inference for Gene Expression and Proteomics*, ed. by K. A. Do, P. Müller, and M. Vannucci, 155–176. Cambridge University Press.

McLachlan, G. J., and D. Peel. 2000. "Mixtures of Factor Analyzers". Chap. 8 in *Finite Mixture Models*, 238–256. John Wiley & Sons, Inc. doi:10.1017/CBO9781107415324.004. arXiv: arXiv:1011.1669v3.

McLachlan, G. J., D. Peel, and R. W. Bean. 2003. "Modelling High-Dimensional Data by Mixtures of Factor Analyzers". *Computational Statistics and Data Analysis* 41 (3-4): 379–388. doi:10.1016/S0167-9473(02)00183-4.

Meng, X.-L., and D. A. van Dyk. 1997. "The EM Algorithm — an Old Folk song Sung to a Fast New Tune (with Discussion)". *Journal of the Royal Statistical Society. Series B* 59 (3): 511–567. doi:10.1111/1467-9868.00082.

— . 1999. "Seeking Efficient Data Augmentation Schemes via Conditional and Marginal Augmentation". *Biometrika* 86 (2): 301–320.

Neal, R. M. 2000. "Markov Chain Sampling Methods for Dirichlet Process Mixture Models". *Journal of Computational and Graphical Statistics* 9 (2): 249–265.

Paisley, J., and L. Carin. 2009. "Nonparametric Factor Analysis with Beta Process Priors". In *Proceedings of the 26th International Conference on Machine Learning*, 1–8. Montreal, Canada: ACM Press. doi:10.1145/1553374.1553474.

Papaspiliopoulos, O., and G. O. Roberts. 2008. "Retrospective Markov Chain Monte Carlo Methods for Dirichlet Process Hierarchical Models". *Biometrika* 95 (1): 169–186. doi:10.1093/biomet/asm086.

Quintana, F. A., and P. Müller. 2004. "Nonparametric Bayesian Data Analysis". *Statistical Science* 19 (1): 95–110. doi:10.1214/088342304000000017.

Reiersøl, O. 1950. "On the Identifiability of Parameters in Thurstone's Multiple Factor Analysis". *Psychometrika* 15 (2): 121–149.

Scott, S. L. 2011. "Data Augmentation, Frequentist Estimation, and the Bayesian Analysis of Multinomial Logit Models". *Statistical Papers* 52 (1): 87–109. doi:10.1007/s00362-009-0205-0.

Sethuraman, J. 1994. "A Constructive Definition of Dirichlet Priors". *Statistica Sinica* 4 (2): 639–650.

Thurstone, L. L. 1934. "The Vectors of Mind". *The Psychological Review* (Chicago) 41 (1): 1–32.

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.

Williams, B. 2015. "Identification of the Linear Factor Model". *Working Paper*.

Xiyun, J., 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.

Yang, M., D. B. Dunson, and D. Baird. 2010. "Semiparametric Bayes Hierarchical Models with Mean and Variance Constraints". *Computational Statistics & Data Analysis* 54 (9): 2172–2186. doi:10.1016/j.csda.2010.03.025.

Yau, C., O. Papaspiliopoulos, G. O. Roberts, and C. C. Holmes. 2011. "Bayesian Non-Parametric Hidden Markov Models with Applications in Genomics". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73 (1): 37–57.

Zhang, X., W. J. Boscardin, and T. R. Belin. 2006. "Sampling Correlation Matrices in Bayesian Models With Correlated Latent Variables". *Journal of Computational and Graphical Statistics* 15 (4): 880–896. doi:10.1198/106186006X160050.

A Prior Distribution

[REMI: I have revised the whole appendices to make them consistent for both finite and infinite cases.]

Proposition 2. *Consider the parameters $\widetilde{\mu}$ and $\widetilde{D}$ defined in eqs. (7) and (8) and the mappings from $\widetilde{\vartheta}$ to ϑ as defined in eqs. (9) and (10). Then, the normal-inverse-Wishart prior distribution specified on $\widetilde{\vartheta}_k = \{\widetilde{\mu}_k, \widetilde{\Phi}_k\}$ in eqs. (3) and (4), for $k \in \mathcal{K}$, where in the finite case $\mathcal{K} = \{1, \ldots, K\}$ and in the infinite case $\mathcal{K} = \mathcal{I}^{(al)}$, implies that*

$$
\begin{aligned}
f(\vartheta \,|\nu_0, A_0, \beta) \propto & \\
& \left|\sum_{k\in\mathcal{K}} \Phi_k^{-1}\right|^{-\frac{1}{2}} \left|\prod_{k\in\mathcal{K}} \Phi_k\right|^{-\frac{\nu_0+P+2}{2}} \left(\prod_{j=1}^{P}\sum_{k\in\mathcal{K}} \left[\Phi_k^{-1}\right]_{[jj]}\right)^{-|\mathcal{K}|\nu_0} \\
& \times \exp\left\{-\frac{1}{2A_0}\left(\sum_{k\in\mathcal{K}} \mu_k'\Phi_k^{-1}\mu_k - \left(\sum_{k\in\mathcal{K}} \Phi_k^{-1}\mu_k\right)'\left(\sum_{k\in\mathcal{K}} \Phi_k^{-1}\right)^{-1}\left(\sum_{k\in\mathcal{K}} \Phi_k^{-1}\mu_k\right)\right)\right\} \\
& \times \mathbf{1}\{\mu = 0\}\,\mathbf{1}\{\mathrm{diag}(\Phi) = \iota_P\},
\end{aligned}
\tag{A1}
$$

where $[\cdot]_{[jj]}$ denotes the j^{th} diagonal element of the corresponding matrix, $|\mathcal{K}|$ is the cardinal number of the set $\mathcal{K}$, and the conditions in the two indicator functions in the last line enforce the constraints on the location and scale of the latent factors, see eqs. (11) and (12).

Note that the dependence on the mixture weights $\beta = (\beta_1, \ldots, \beta_K)$ is hidden in the constraints imposed via the indicator functions. Also, note that this density does not depend on the scaling parameter s_0 of the inverse-Wishart distribution of the covariance matrices in the auxiliary model. This parameter controls the degree of inflation of the parameters in the augmented model, but has no influence on the prior distribution of the parameters in the identified model.

A.1 Proof of propositions 1 and 2

[OM: Remi also arrange the proof appropriately and please do not use the word "prior" confusingly]

Induced joint prior distribution. The joint distribution of $\{\widetilde{\mu}, \widetilde{D}, \mu_{\mathcal{K}}, \Phi_{\mathcal{K}}\}$ is derived from the distribution of $\{\widetilde{\mu}_{\mathcal{K}}, \widetilde{\Phi}_{\mathcal{K}}\}$ in the expanded model using a transformation of random variables. The restrictions in eqs. (11) and (12) imply that:

$$
f(\widetilde{\mu}, \widetilde{D}, \mu_{\mathcal{K}}, \Phi_{\mathcal{K}}) = f(\widetilde{\mu}, \widetilde{D}, \mu_{-k}, \Phi_{-k}, \Phi_k^L) \underbrace{f(\mu_k \mid \widetilde{\mu}, \widetilde{D}, \mu_{-k}, \Phi_{\mathcal{K}})}_{\mathbf{1}\{\mu=0_P\}} \underbrace{f(\Phi_k^D \mid \widetilde{\mu}, \widetilde{D}, \mu_{-k}, \Phi_{-k}, \Phi_k^L)}_{\mathbf{1}\{\mathrm{diag}(\Phi)=\iota_P\}}
\tag{A2}
$$

where $\mu_{\mathcal{K}} = \{\mu_k\}_{k\in\mathcal{K}}$ and $\Phi_{\mathcal{K}} = \{\Phi_k\}_{k\in\mathcal{K}}$, and Φ_k^D and Φ_k^L denote, respectively, the diagonal elements and the lower triangular part (excluding the diagonal elements) of Φ_k. The first density is obtained from the transformation of random variables $(\widetilde{\mu}_{\mathcal{K}}, \widetilde{\Phi}_{\mathcal{K}}) \rightarrow (\widetilde{\mu}, \widetilde{D}, \mu_{-k}, \Phi_{-k}, \Phi_k^L)$, such that:

$$f(\widetilde{\mu}, \widetilde{D}, \mu_{-k}, \Phi_{-k}, \Phi_k^L) = f(\widetilde{\mu}_{\mathcal{K}}, \widetilde{\Phi}_{\mathcal{K}})\, \mathcal{J}\{(\widetilde{\mu}_{\mathcal{K}}, \widetilde{\Phi}_{\mathcal{K}}) \rightarrow (\widetilde{\mu}, \widetilde{D}, \mu_{-k}, \Phi_{-k}, \Phi_k^L)\}, \tag{A3}$$

where $\mathcal{J}\{(\cdot) \rightarrow (\cdot)\}$ is the Jacobian of the corresponding transformation.

Since the mixture parameters $\{\widetilde{\mu}_k, \widetilde{\Phi}_k\}$ are assumed to be independent across mixture components and to follow a normal-inverse-Wishart distribution for each $k \in \mathcal{K}$ in the expanded model, see eqs. (3) and (4), the joint distribution of the corresponding parameters in the identified model $\mu_{\mathcal{K}}$ and $\Phi_{\mathcal{K}}$ and of the working parameters $\widetilde{\mu}$ and $\widetilde{D}$ is derived from eqs. (A2) and (A3) as follows, and without loss of generality:[12]

$$\begin{aligned}
f(\widetilde{\mu}, \widetilde{D}, \mu_{\mathcal{K}}, \Phi_{\mathcal{K}}) &= f(\widetilde{\mu}_{\mathcal{K}}, \widetilde{\Phi}_{\mathcal{K}})\, \mathcal{J}\{(\widetilde{\mu}_{\mathcal{K}}, \widetilde{\Phi}_{\mathcal{K}}) \rightarrow (\widetilde{\mu}, \widetilde{D}, \mu_{-k}, \Phi_{-k}, \Phi_k^L)\} \\
&\quad \times \mathbf{1}\{\mu = 0_P\}\mathbf{1}\{\mathrm{diag}(\Phi) = \iota_P\}, \\
&\propto \prod_{k\in\mathcal{K}} \left|\, \widetilde{\Phi}_k \,\right|^{-1/2} \exp\left\{-\frac{1}{2A_0}\widetilde{\mu}_k'\widetilde{\Phi}_k^{-1}\widetilde{\mu}_k\right\} \left|\, \widetilde{\Phi}_k \,\right|^{-\frac{\nu_0+P+1}{2}} \exp\left\{-\frac{s_0}{2}\mathrm{tr}\left(\widetilde{\Phi}_k^{-1}\right)\right\} \\
&\quad \prod_{i=1}^{P} \left(\widetilde{D}_i\right)^{\frac{|\mathcal{K}|(P+2)-3}{2}} \mathbf{1}\{\mu = 0_P\}\mathbf{1}\{\mathrm{diag}(\Phi) = \iota_P\}, \\
&\propto \exp\left\{-\frac{1}{2A_0}\left(\widetilde{\mu}'\widetilde{D}^{-\frac{1}{2}}\left(\sum_{k\in\mathcal{K}}\Phi_k^{-1}\right)\widetilde{D}^{-\frac{1}{2}}\widetilde{\mu} + 2\widetilde{\mu}'\widetilde{D}^{-\frac{1}{2}}\left(\sum_{k\in\mathcal{K}}\Phi_k^{-1}\mu_k\right)\right)\right\} \qquad \text{(A4)} \\
&\quad \times \prod_{i=1}^{P} \left(\widetilde{D}_i\right)^{-\frac{|\mathcal{K}|\nu_0+1}{2}-1} \exp\left\{-\frac{s_0}{2}\sum_{k\in\mathcal{K}}\left[\Phi_k^{-1}\right]_{[ii]}\widetilde{D}_i^{-1}\right\} \qquad \text{(A5)} \\
&\quad \times \prod_{k\in\mathcal{K}} \left|\, \Phi_k \,\right|^{-\frac{\nu_0+P+2}{2}} \exp\left\{-\frac{\mu_k'\Phi_k^{-1}\mu_k}{2A_0}\right\} \\
&\quad \times \mathbf{1}\{\mu = 0_P\}\mathbf{1}\{\mathrm{diag}(\Phi) = \iota_P\},
\end{aligned}$$

where the Jacobian of the transformation is proportional to $\prod_{i=1}^{P}\left(\widetilde{D}_i\right)^{\frac{|\mathcal{K}|(P+2)-3}{2}}$, see appendix A.2.

The kernel can be factorized as

$$f(\widetilde{\mu}, \widetilde{D}, \mu_{\mathcal{K}}, \Phi_{\mathcal{K}}) = f(\widetilde{\mu} \mid \widetilde{D}, \mu_{\mathcal{K}}, \Phi_{\mathcal{K}}) f(\widetilde{D} \mid \mu_{\mathcal{K}}, \Phi_{\mathcal{K}}) f(\mu_{\mathcal{K}}, \Phi_{\mathcal{K}}),$$

and the three distributions on the right-hand side can be retrieved as follows. The conditional distribution of $\widetilde{\mu}$ is obtained from eq. (A4), which is the kernel of a Gaussian

[12] Because of the identification constraints on the mixture means and variances, the mean and variance of one mixture component are redundant and can be discarded. Which component is discarded does not affect the results.

distribution:

$$\widetilde{\mu} \mid \widetilde{D}, \mu_{\mathcal{K}}, \Phi_{\mathcal{K}}, A_0 \sim \mathcal{N}\left(-\widetilde{D}^{\frac{1}{2}} \left(\sum_{k \in \mathcal{K}} \Phi_k^{-1} \right)^{-1} \left(\sum_{k \in \mathcal{K}} \Phi_k^{-1} \mu_k \right) ; \; A_0 \widetilde{D}^{\frac{1}{2}} \left(\sum_{k \in \mathcal{K}} \Phi_k^{-1} \right)^{-1} \widetilde{D}^{\frac{1}{2}} \right).$$

The conditional distribution of $\widetilde{D}$ is obtained by integrating out $\widetilde{\mu}$, using the kernel in eq. (A5) and completing the normalizing constant that depends on $\widetilde{D}$ in eq. (A4):

$$\begin{aligned} f(\widetilde{D} \mid \mu_{\mathcal{K}}, \Phi_{\mathcal{K}}, \nu_0, s_0) &\propto \int f(\widetilde{\mu}, \widetilde{D}, \mu_{\mathcal{K}}, \Phi_{\mathcal{K}}) d\widetilde{\mu}, \\ &\propto \left| \widetilde{D} \right|^{\frac{1}{2}} \prod_{i=1}^{P} \left(\widetilde{D}_i \right)^{-\frac{|\mathcal{K}|\nu_0+1}{2}-1} \exp\left\{ -\frac{s_0}{2} \sum_{k \in \mathcal{K}} \left[\Phi_k^{-1} \right]_{[ii]} \widetilde{D}_i^{-1} \right\}, \\ &\propto \prod_{i=1}^{P} \left(\widetilde{D}_i \right)^{-\frac{|\mathcal{K}|\nu_0}{2}-1} \exp\left\{ -\frac{s_0}{2} \sum_{k \in \mathcal{K}} \left[\Phi_k^{-1} \right]_{[ii]} \widetilde{D}_i^{-1} \right\}, \end{aligned}$$

which results in a product of kernels of inverse-Gamma distributions:

$$\widetilde{D}_i \mid \Phi_{\mathcal{K}}, \nu_0, s_0 \sim \mathcal{IG}\left(\frac{|\mathcal{K}|\nu_0}{2}; \; \frac{s_0}{2} \sum_{k \in \mathcal{K}} \left[\Phi_k^{-1} \right]_{[ii]} \right),$$

for $i = 1, \ldots, P$.

Finally, the kernel of the marginal distribution of the mixture parameters in the identified model is obtained by integrating both $\widetilde{\mu}$ and $\widetilde{D}$ out of the joint distribution:

$$f(\mu_{\mathcal{K}}, \Phi_{\mathcal{K}} \mid A_0, \nu_0) \propto \iint f(\widetilde{\mu}, \mu_{\mathcal{K}}, \widetilde{D}, \Phi_{\mathcal{K}}) \, d\widetilde{\mu} \, d\widetilde{D},$$

which produces the kernel in eq. (A1).

□

A.2 Jacobian of the transformation

The Jacobian corresponding to the change of variables that allows to move from the expanded model to the identified model can be derived in several steps. Because of the restrictions on the parameters of the identified model ($\mu = 0_P$ and $\text{diag}(\Phi) = \iota_P$), one of the mixture means and the diagonal elements of one of the covariance matrices are redundant in the parameter transformation and can be left aside in the derivation. The subscript $-k$ indicates that the k^{th} element of the corresponding set is left out, e.g., $\mu_{-k} = \{\mu_j \mid j \in \mathcal{K}, j \neq k\}$. We denote Φ_k^L the lower triangular elements of Φ_k, excluding the diagonal elements. Without loss of generality, we derive the Jacobian for the case where the mean and the diagonal elements of the covariance matrix of the kth mixture

component are left aside:

$$
\begin{aligned}
&\mathcal{J}\{(\widetilde{\mu}_{\mathcal{K}},\widetilde{\Phi}_{\mathcal{K}}) \to (\widetilde{\mu}, \widetilde{D}, \mu_{-k}, \Phi_{-k}, \Phi_k^L)\} \\
&\quad = \mathcal{J}\{(\widetilde{\mu}_{\mathcal{K}}, \widetilde{\Phi}_{\mathcal{K}}) \to (\widetilde{\mu}, \widetilde{\mu}_{-k}, \widetilde{\Phi}_{\mathcal{K}})\} \\
&\qquad \times \mathcal{J}\{(\widetilde{\mu}, \widetilde{\mu}_{-k}, \widetilde{\Phi}_{\mathcal{K}}) \to (\widetilde{\mu}, \widetilde{\mu}_{-k}, \widetilde{\Phi}, \widetilde{\Phi}_{-k})\} \\
&\qquad \times \mathcal{J}\{(\widetilde{\mu}, \widetilde{\mu}_{-k}, \widetilde{\Phi}, \widetilde{\Phi}_{-k}) \to (\widetilde{\mu}, \widetilde{\mu}_{-k}, \widetilde{D}, \Phi, \widetilde{\Phi}_{-k})\} \\
&\qquad \times \mathcal{J}\{(\widetilde{\mu}, \widetilde{\mu}_{-k}, \widetilde{D}, \Phi, \widetilde{\Phi}_{-k}) \to (\widetilde{\mu}, \mu_{-k}, \widetilde{D}, \Phi, \widetilde{\Phi}_{-k})\} \\
&\qquad \times \mathcal{J}\{(\widetilde{\mu}, \mu_{-k}, \widetilde{D}, \Phi, \widetilde{\Phi}_{-k}) \to (\widetilde{\mu}, \mu_{-k}, \widetilde{D}, \Phi, \Phi_{-k})\} \\
&\qquad \times \mathcal{J}\{(\widetilde{\mu}, \mu_{-k}, \widetilde{D}, \Phi, \Phi_{-k}) \to (\widetilde{\mu}, \mu_{-k}, \widetilde{D}, \Phi_{-k}, \Phi_k^L)\} \\
&\quad = \left(\frac{1}{p_k}\right)^P \times \left(\frac{1}{p_k}\right)^{\frac{P(P+1)}{2}} \times \prod_{i=1}^P \left(\widetilde{D}_i\right)^{\frac{P-1}{2}} \times \prod_{i=1}^P \left(\widetilde{D}_i\right)^{\frac{|\mathcal{K}|-1}{2}} \\
&\qquad \times \prod_{i=1}^P \left(\widetilde{D}_i\right)^{\frac{(P+1)(|\mathcal{K}|-1)}{2}} \times p_k^{\frac{P(P-1)}{2}}, \\
&\quad = \left(\frac{1}{p_k}\right)^{2P} \prod_{i=1}^P \left(\widetilde{D}_i\right)^{\frac{|\mathcal{K}|(P+2)-3}{2}}, \\
&\quad \propto \prod_{i=1}^P \left(\widetilde{D}_i\right)^{\frac{|\mathcal{K}|(P+2)-3}{2}},
\end{aligned}
\tag{A6}
$$

where the Jacobian in line A6 is derived as in Zhang et al. (2006).

B Details on MCMC Sampler

This appendix provides technical details on the MCMC sampler. These steps are presented in their generic form and can be used both for the finite and the infinite cases, where in the former $\mathcal{K} = \{1, \ldots, K\}$, while in the latter $\mathcal{K} = \mathcal{I}^{(\mathrm{al})}$, respectively, and $|\mathcal{K}|$ is the cardinal number of $\mathcal{K}$.

B.1 Sampling the idiosyncratic variances (step 1)

The inverse-Gamma prior in eq. (15) provides the following posterior, for $q = 1, \ldots, Q$:

$$
\sigma_q^2 \mid Y, \theta, \delta, \Lambda, a_0, b_0 \sim \mathcal{IG}\left(a_0 + \frac{N}{2};\, b_0 + \frac{1}{2}\sum_{i=1}^N (Y_i - \delta - \Lambda\theta_i)^2\right).
$$

B.2 Sampling the latent factors (step 2b)

For each $i = 1, \ldots, N$, draw $\theta_i \mid Y_i, G_i, \vartheta, \delta, \Lambda, \Sigma \sim \mathcal{N}(B_{\theta_i} b_{\theta_i};\, B_{\theta_i})$, where

$$
B_{\theta_i}^{-1} = \Lambda' \Sigma^{-1} \Lambda + \Phi_{G_i}^{-1}, \qquad b_{\theta_i} = \Lambda' \Sigma^{-1} (Y_i - \delta) + \Phi_{G_i}^{-1} \mu_{G_i}.
$$

B.3 Sampling the working parameters conditional on the latent factors in the expanded model (step 3a)

The joint conditional distribution of the working parameters, given their prior distributions expressed in eqs. (18) and (19), is proportional to:

$$
\begin{aligned}
p(\widetilde{\mu}, \widetilde{D} \mid \widetilde{\theta}, G, \vartheta) &\propto p(\widetilde{\theta} \mid \widetilde{\mu}, \widetilde{D}, G, \vartheta)\, p(\widetilde{\mu} \mid \widetilde{D}, \vartheta)\, p(\widetilde{D} \mid \vartheta), \\
&\propto \exp\left\{ -\frac{1}{2} \left[\widetilde{\mu}' \widetilde{D}^{-\frac{1}{2}} \left(\sum_{k \in \mathcal{K}} (N_k + A_0^{-1}) \Phi_k^{-1} \right) \widetilde{D}^{-\frac{1}{2}} \widetilde{\mu} \right.\right. \\
&\qquad \left.\left. -2 \widetilde{\mu}' \widetilde{D}^{-\frac{1}{2}} \sum_{k \in \mathcal{K}} \Phi_k^{-1} \left(\left[\widetilde{D}^{-\frac{1}{2}} \sum_{i \in \mathcal{I}_k} \widetilde{\theta}_i \right] - (N_k + A_0^{-1}) \mu_k \right) \right] \right\} \\
&\quad \times \left| \widetilde{D} \right|^{-\frac{|\mathcal{K}|\nu_0 + N + 1}{2} - 1} \exp\left\{ -\frac{1}{2} \sum_{k \in \mathcal{K}} \operatorname{tr} \left(\widetilde{D}^{-\frac{1}{2}} \Phi_k^{-1} \widetilde{D}^{-\frac{1}{2}} \left[\sum_{i \in \mathcal{I}_k} \widetilde{\theta}_i \widetilde{\theta}_i' + s_0 I_P \right] \right) \right. \\
&\qquad \left. + \sum_{k \in \mathcal{K}} \mu_k' \Phi_k^{-1} \widetilde{D}^{-\frac{1}{2}} \sum_{i \in \mathcal{I}_k} \widetilde{\theta}_i \right\}.
\end{aligned} \tag{B1}
$$

This provides the kernel of a normal distribution for $\widetilde{\mu}$ conditional on $\widetilde{D}$ and on the remaining parameters:

$$
\widetilde{\mu} \mid \widetilde{\theta}, \widetilde{D}, G, \vartheta \sim \mathcal{N}\left(\widetilde{D}^{\frac{1}{2}} B_2 (B_1(\widetilde{D}) - B_3);\, \widetilde{D}^{\frac{1}{2}} B_2 \widetilde{D}^{\frac{1}{2}} \right),
$$

with:

$$
B_1(\widetilde{D}) = \sum_{k \in \mathcal{K}} \Phi_k^{-1} \widetilde{D}^{-\frac{1}{2}} \sum_{i \in \mathcal{I}_k} \widetilde{\theta}_i, \qquad B_2^{-1} = \sum_{k \in \mathcal{K}} (N_k + A_0^{-1}) \Phi_k^{-1},
$$

$$
B_3 = \sum_{k \in \mathcal{K}} (N_k + A_0^{-1}) \Phi_k^{-1} \mu_k.
$$

As for the other working parameters $\widetilde{D}$, the kernel of their conditional distribution is obtained by integrating $\widetilde{\mu}$ out of the joint distribution, by completing the normalizing

constant of eq. (B1):

$$
\begin{aligned}
p(\widetilde{D} \mid \widetilde{\theta}, G, \vartheta) &= \int p(\widetilde{\mu}, \widetilde{D} \mid \widetilde{\theta}, G, \vartheta) d\widetilde{\mu}, \\
&\propto \left| \widetilde{D} \right|^{-\frac{|\mathcal{K}|\nu_0 + N}{2} - 1} \exp \left\{ \frac{1}{2} B_1(\widetilde{D})' B_2 \left(B_1(\widetilde{D}) - 2B_3 \right) \right. \\
&\qquad - \frac{1}{2} \sum_{k \in \mathcal{K}} \operatorname{tr} \left(\widetilde{D}^{-\frac{1}{2}} \Phi_k^{-1} \widetilde{D}^{-\frac{1}{2}} \left[\sum_{i \in \mathcal{I}_k} \widetilde{\theta}_i \widetilde{\theta}_i' + s_0 I_P \right] \right) \\
&\qquad \left. + \sum_{k \in \mathcal{K}} \mu_k' \Phi_k^{-1} \widetilde{D}^{-\frac{1}{2}} \sum_{i \in \mathcal{I}_k} \widetilde{\theta}_i \right\},
\end{aligned}
$$

which is not the kernel of a known distribution. However, $\widetilde{D}$ can be simulated with a Metropolis-Hastings step.

Metropolis-Hastings step to sample $\widetilde{D}$. As proposal distribution for each of the diagonal elements $j = 1, \ldots, P$ of $\widetilde{D}$, a log-normal distribution is used, parametrized such that its mode is equal to $\widetilde{D}_j$:

$$
\begin{aligned}
\widetilde{D}_j^\star \mid (\widetilde{D}_j, \rho^2) &\sim \ln \mathcal{N} \left(\ln \widetilde{D}_j + \rho^2;\, \rho^2 \right), \\
q(\widetilde{D}^\star \mid \widetilde{D}, \rho^2) &\propto \prod_{j=1}^{P} \frac{1}{\widetilde{D}_j} \exp \left\{ -\frac{1}{2\rho^2} \left(\ln \widetilde{D}_j^\star - \ln \widetilde{D}_j - \rho^2 \right)^2 \right\}.
\end{aligned}
$$

The P proposed values $\widetilde{D}^\star$ are accepted as new draws for $\widetilde{D}$ with probability:

$$
\alpha(\widetilde{D}^\star \mid \widetilde{D}) = \min \left\{ 1; \frac{f(\widetilde{D}^\star \mid \widetilde{\theta}, G, \vartheta)}{f(\widetilde{D} \mid \widetilde{\theta}, G, \vartheta)} \frac{q(\widetilde{D} \mid \widetilde{D}^\star, \rho^2)}{q(\widetilde{D}^\star \mid \widetilde{D}, \rho^2)} \right\},
$$

where the second ratio simplifies to $\ln \frac{q(\widetilde{D} | \widetilde{D}^\star, \rho^2)}{q(\widetilde{D}^\star | \widetilde{D}, \rho^2)} = \sum_{j=1}^{P} (\ln \widetilde{D}_j - \ln \widetilde{D}_j^\star)$. The parameter ρ^2 is a tuning parameter that influences the acceptance rate of the Metropolis-Hastings algorithm. We use $\rho^2 = 1/N$ in our applications, which provides an acceptance rate around 60%.

B.4 Sampling the intercept terms and factor loadings (step 3b-c)

In the expanded model, the prior distributions specified in eqs. (20) and (21) result in the following posteriors, for each manifest variable $q = 1, \ldots, Q$:

$$
\begin{aligned}
\widetilde{\Lambda}_q \mid Y_q, \widetilde{\theta}, \sigma_q^2, \widetilde{\mu}, \widetilde{D} &\sim \mathcal{N} \left(B_{\widetilde{\Lambda}_q} b_{\widetilde{\Lambda}_q};\, B_{\widetilde{\Lambda}_q} \right), \\
\widetilde{\delta}_q \mid Y_q, \widetilde{\theta}, \widetilde{\Lambda}_q, \sigma_q^2, \widetilde{\mu}, \widetilde{D} &\sim \mathcal{N} \left(B_{\widetilde{\delta}_q} b_{\widetilde{\delta}_q};\, B_{\widetilde{\delta}_q} \right),
\end{aligned}
$$

with:

$$B_{\widetilde{\delta}_q}^{-1} = c_0 + \frac{N}{\sigma_q^2}, \qquad b_{\widetilde{\delta}_q} = \frac{1}{\sigma_q^2}\sum_{i=1}^{N}\left(Y_{qi} - \widetilde{\Lambda}_q'\widetilde{\theta}_i\right) - \frac{\widetilde{\Lambda}_q'\widetilde{\mu}}{c_0},$$

$$B_{\widetilde{\Lambda}_q}^{-1} = \frac{\widetilde{\theta}'\widetilde{\theta}}{\sigma_q^2} + \frac{\widetilde{\mu}\widetilde{\mu}'}{c_0} + \frac{\widetilde{D}}{d_0} - B_{\widetilde{\delta}_q} b_q b_q', \qquad b_{\widetilde{\Lambda}_q} = \frac{1}{\sigma_q^2}\left(\widetilde{\theta}' Y_q - b_q B_{\widetilde{\delta}_q}\left(\sum_{i=1}^{N} Y_{qi}\right)\right),$$

where $b_q = c_0^{-1}\widetilde{\mu} + \frac{1}{\sigma_q^2}\sum_{i=1}^{N}\widetilde{\theta}_i$.

B.5 Sampling the parameters of the mixture components in the expanded model (step 4)

The conjugate normal-inverse-Wishart prior distribution specified on the mixture parameters in eqs. (3) and (4) results in the following posterior distribution for the non-empty mixture components $k \in \mathcal{I}^{(\mathrm{al})}$:

$$\widetilde{\Phi}_k \mid \widetilde{\theta}, G \sim \mathcal{IW}\left(\nu_0 + N_k;\; s_0 I_P + \sum_{i\in\mathcal{I}_k}\widetilde{\theta}_i\widetilde{\theta}_i' - \frac{\left(\sum_{i\in\mathcal{I}_k}\widetilde{\theta}_i\right)\left(\sum_{i\in\mathcal{I}_k}\widetilde{\theta}_i\right)'}{N_k + A_0^{-1}}\right),$$

$$\widetilde{\mu}_k \mid \widetilde{\Phi}_k, \widetilde{\theta}, G \sim \mathcal{N}\left(\frac{\sum_{i\in\mathcal{I}_k}\widetilde{\theta}_i}{N_k + A_0^{-1}};\; \frac{\widetilde{\Phi}_k}{N_k + A_0^{-1}}\right).$$

The parameters of the empty mixture components ("dead" components) are sampled from their prior distribution. This is straightforward to do in the finite mixture case, but infeasible in the infinite case. Instead, we sample them later and retrospectively when new mixture components are required in steps 5 and 6, see appendix B.6.

B.6 Sampling the mixture group indicators and the mixture probabilities (steps 5 and 6)

In the finite case, each observation $i = 1, \ldots, N$ is allocated to mixture group k with probability $p\left(G_i = k \mid \widetilde{\theta}_i, p_k, \widetilde{\vartheta}_k\right) \propto p_k \left|\widetilde{\Phi}_k\right|^{-\frac{1}{2}} \phi_P\left(\widetilde{\Phi}_k^{-\frac{1}{2}}(\widetilde{\theta}_i - \widetilde{\mu}_k)\right)$ where $\phi_P(\cdot)$ denotes the probability density function of the multivariate standard normal distribution. The random variables underlying the stick-breaking process are updated as $V_k \mid G, \alpha \sim \mathcal{B}eta\left(N_k + 1; \alpha + \sum_{j=k+1}^{K} N_j\right)$, for $k = 1, \ldots, K-1$. Taking $V_K = 1$, the corresponding mixture probabilities are computed as in eq. (5).

In the infinite case, we implement Algorithm 2 of Papaspiliopoulos and Roberts (2008, p. 176) to introduce new mixture components into the model on the fly. The parameters of the corresponding new mixture components are sampled from their prior distribution *retrospectively* as they become required. [REMI: Not sure how far into details we should go here. This is a plain application of your algorithm.]

B.7 Sampling the concentration parameter α (step 7)

Following Escobar and West (1995), the Gamma prior distribution specified on α in eq. (16) results in a posterior that is a mixture of two Gamma distributions:

$$\eta \mid \alpha, K^+ \sim \mathcal{B}eta(\alpha + 1; N),$$
$$\alpha \mid \eta, K^+ \sim \pi_\eta \mathcal{G}\big(g_0 + K^+;\, h_0 - \log(\eta)\big) + (1 - \pi_\eta)\mathcal{G}\big(g_0 + K^+ - 1;\, h_0 - \log(\eta)\big)\,,$$

with $\pi_\eta/(1 - \pi_\eta) = (g_0 + K^+ - 1)/(N(h_0 - \log(\eta)))$, and where K^+ denotes the number of non-empty mixture components.