Bayesian Exploratory Factor Analysis

by

Gabriella CONTI
Sylvia FRÜHWIRTH-SCHNATTER*)
James J. HECKMAN
Rémi PIATEK

Working Paper No. 1408

June 2014

Supported by the Austrian Science Funds

FWF

The Austrian Center for Labor Economics and the Analysis of the Welfare State

JKU Linz
Department of Economics
Altenberger Strasse 69
4040 Linz, Austria
www.labornrn.at

Corresponding author: sylvia.fruehwirth-schnatter@wu.ac.at
Phone: +43-1-313 36-5581

Bayesian Exploratory Factor Analysis*

Gabriella Conti[1], Sylvia Frühwirth-Schnatter[2], James J. Heckman[3,4], and Rémi Piatek†[5]

[1]Department of Applied Health Research, University College London, UK
[2]Vienna University of Economics and Business, Austria
[3]Department of Economics, University of Chicago, USA
[4]American Bar Foundation, USA
[5]Department of Economics, University of Copenhagen, Denmark

Abstract

This paper develops and applies a Bayesian approach to Exploratory Factor Analysis that improves on *ad hoc* classical approaches. Our framework relies on dedicated factor models and simultaneously determines the number of factors, the allocation of each measurement to a unique factor, and the corresponding factor loadings. Classical identification criteria are applied and integrated into our Bayesian procedure to generate models that are stable and clearly interpretable. A Monte Carlo study confirms the validity of the approach. The method is used to produce interpretable low dimensional aggregates from a high dimensional set of psychological measurements.

JEL Codes: C11; C38; C63.

Keywords: Bayesian Factor Models; Exploratory Factor Analysis; Identifiability; Marginal Data Augmentation; Model Expansion; Model Selection.

*This paper is forthcoming in *Journal of Econometrics*. We thank the editor, John Geweke, and two anonymous referees for comments. This work was presented at the third European Seminar on Bayesian Econometrics (November 2012, Vienna). We thank the participants for their helpful comments, and especially Xiao-Li Meng and our discussant Jesús Crespo Cuaresma. We also gratefully acknowledge support from NIH R01 HD054702 and R37 HD065072, the American Bar Foundation, The J.B. and M.K. Pritzker Foundation, the Geary Institute at University College Dublin, a grant from the European Research Council DEVHEALTH-269874, and an anonymous funder. The research of the second author was partly funded by the Austrian Science Fund (FWF): S10309-G16. The views expressed in this paper are those of the authors and not necessarily those of the funders. A Web Appendix containing additional material is available at http://heckman.uchicago.edu/BayesFA.

†Corresponding author: Rémi Piatek. Address: Department of Economics, University of Copenhagen, Øster Farimagsgade 5, Building 26, DK-1353 Copenhagen K, Denmark. Email: Remi.Piatek@econ.ku.dk. Tel. (+45) 35 32 30 35.

1 Introduction

As the production of social statistics proliferates, aggregation and condensation of data have become increasingly important. William Barnett has made and continues to make numerous important contributions to constructing economically meaningful monetary aggregates (see, e.g., Barnett and Chauvet, 2011). In the spirit of Barnett's pioneering research, this paper addresses the problem of constructing reliable and interpretable aggregates from myriad measures. It is the first paper in the literature on Bayesian factor analysis to make inference on a model where all measurements load onto at most one factor, and factors are correlated. The model allows for the dimension of the latent structure to be unknown *a priori*, and the allocation of measurements to factors is part of the inference procedure. Classical identification criteria are invoked and applied to the analysis to generate interpretable posterior distributions.

The abundance of measures is both an opportunity and a challenge in many empirical applications. The main question—both from a methodological and an applied standpoint—is how to condense the available information into interpretable aggregates. Thurstone (1934) postulated criteria and developed analytical methods for estimating and identifying factor models with *perfect simple structure*, where each measurement is related to at most one latent factor. In his view, models with simple structure were transparent and easily interpreted. He developed the method of "oblique" factor analysis by arguing that correlated factors were a more plausible representation of reality (Thurstone, 1947). Cattell (1952, 1966); Carroll (1953); Saunders (1953); Ferguson (1954) and Hofmann (1978) are major exponents of the concept of parsimony in the Thurstone tradition. We call Thurstone's simple structure a *dedicated structure* in this paper. It dedicates all measures to at most one factor. This representation is widely used in economics (Heckman et al., 2006; Cunha et al., 2010; Conti et al., 2010; Baron and Cobb-Clark, 2010).

Exploratory Factor Analysis is a well developed classical procedure for doing dedicated factor analysis (Gorsuch, 1983, 2003). The various steps required in executing classical Exploratory Factor Analyses (EFA) are all subject to a certain degree of arbitrariness and entail *ad hoc* judgments. Classical EFA proceeds in four separate steps: (i) selecting the dimension of the factor model; (ii) allocating measurements to factors; (iii) estimating factor loadings; and (iv) discarding measurements that load on multiple factors. A variety of methods are available to select the dimension of the latent structure, to extract and rotate factors (Gorsuch, 2003; Costello and Osborne, 2005; Jennrich, 2001, 2002, 2004, 2006, 2007). Our empirical analysis shows that each of the choices made by analysts at the various stages of a classical EFA has substantial consequences on the estimated factor structure.

This paper develops an integrated Bayesian approach to EFA that simultaneously selects the dimension of the factor model, the allocation of measurements to factors, and the factor loadings. Our method uses all of the available information by *not* discarding measurements besides those that do not load on any factors. The procedure is justified by the usual appeal to the optimality of Bayes procedures (see Berger, 1985). Different from the classical literature in EFA, in our approach the number of factors is not determined in a first step, but inferred along with other parameters. Our work advances the Bayesian approach to factor analysis, because of the attention paid to the identification of the model. One of our main contributions is to incorporate classical identification criteria into a Bayesian inference procedure. In so doing, we are able to generate posterior distributions that are stable and models that are clearly interpretable. The identifiability of the model is a key feature of the algorithm. In this respect, our paper bridges a gap between the classical and the Bayesian literatures.

Most articles on Bayesian factor analysis rely on a lower-triangular specification for the factor loading matrix to achieve identification (West, 2003; Lopes and West, 2004; Lucas et al., 2006; Carvalho et al., 2008). This approach, first suggested by Anderson and Rubin (1956), has been widely applied (see, for example, Geweke and Zhou, 1996; Aguilar and West, 2000; Carneiro et al., 2003). It achieves identification in the general case, but at the price of *ad hoc* decisions that result in a loss of flexibility—e.g., the choice and the ordering of the measurements at the top of the factor loading matrix is not innocuous. In the framework of sparse factor modelling, the problem becomes more complex, as the structure of the factor loading matrix—in terms of position of the zero elements—is part of the inference problem. Besides the upper triangle of the loading matrix that is fixed to zero *a priori*, the remaining elements in the lower part of the matrix are also allowed to become equal to zero. This introduces new challenges for identification, and additional identifying restrictions are required. Our paper discusses this issue that has, to the best of our knowledge, been overlooked in the literature so far. To tackle this problem, we take a different avenue and incorporate identifying criteria into the prior distribution of model parameters instead of imposing zero restrictions on the factor loading matrix *a priori* (Frühwirth-Schnatter and Lopes, 2012, adopt a related approach).

In the field of Bayesian nonparametrics and machine learning, a strand of literature is dedicated to the inference of factor models with a sparse structure of unknown dimension (Knowles and Ghahramani, 2007; Paisley and Carin, 2009; Bhattacharya and Dunson, 2011), and in a dynamic context with an unknown number of time-dependent factors (Chen et al., 2011). These methods, however, focus on covariance structures, variable selection, or prediction, and identification is not strictly required to achieve these goals from a Bayesian

perspective. No paper in the Bayesian nonparametric literature imposes identifying restrictions on models in its inference algorithm.

Most existing approaches assume uncorrelated factors. Our method is the first in the Bayesian literature to allow for correlated factors in the framework of a model where identification is secured. The specification of correlated factors, combined with the need to produce identified models in a dimension-varying framework, raises challenges for the design of a practical and efficient algorithm that are addressed in this paper.

The paper is organized in the following way. Section 2 presents our framework, which allows for both continuous and binary measurements. We discuss the identification challenges at stake, provide conditions for identification, and explain the constraints they impose on the model. We also introduce the prior specification we adopt to conduct Bayesian inference. Section 3 derives a new Bayesian computational procedure for identifying the latent structure of the model and selecting factors. Section 4 presents a Monte Carlo study that supports the validity of the method. An empirical analysis demonstrates how our method can be applied, and how it uses the information available in the data in comparison with classical EFA. Section 5 concludes.

2 The Model

This section introduces our model, the identification conditions for the model and the prior specification. We develop classical identification conditions for a dedicated factor model. Under standard regularity conditions, satisfaction of classical identification conditions guarantees convergence of the model parameters to asymptotically normal distributions and thus has a large sample justification in addition to a Bayesian justification (Le Cam, 1986). Thus we bridge the two approaches.

2.1 A Dedicated Factor Model with Continuous and Binary Measurements

Consider a set of M continuous and binary measurements arrayed in vector $Y_i = (Y_{i1}, \ldots, Y_{iM})'$ for individual i, $i = 1, \ldots, N$, and matrix $\boldsymbol{Y} = (Y_1, \ldots, Y_N)'$ for the whole sample. To accommodate both types of variables, each measurement is assumed to be determined by an underlying continuous latent variable $Y_{im}^\star$:

$$Y_{im} = \begin{cases} Y_{im}^\star, & \text{if } Y_{im} \text{ is continuous,} \\ \mathbf{1}[Y_{im}^\star > 0], & \text{if } Y_{im} \text{ is binary,} \end{cases}$$

for $m = 1, \ldots, M$.[1] The resulting vector of latent variables $Y_i^\star = (Y_{i1}^\star, \ldots, Y_{iM}^\star)'$ is specified as a function of a set of Q observed variables X_i and K latent factors $\theta_i = (\theta_{i1}, \ldots, \theta_{iK})'$:

$$\underset{(M\times 1)}{Y_i^\star} = \underset{(M\times Q)(Q\times 1)}{\boldsymbol{\beta}\ X_i} + \underset{(M\times K)(K\times 1)}{\boldsymbol{\alpha}\ \theta_i} + \underset{(M\times 1)}{\varepsilon_i}, \tag{1}$$

where the matrix of regression coefficients $\boldsymbol{\beta}$ captures the effect of the covariates on the latent variables, denoted $\boldsymbol{X} = (X_1, \ldots, X_N)'$ and $\boldsymbol{Y}^\star = (Y_1^\star, \ldots, Y_N^\star)'$ respectively. The correlation between the measurements conditional on X_i arises from the factors with loadings $\boldsymbol{\alpha}$. The residual idiosyncratic terms ("uniquenesses") are denoted $\varepsilon_i = (\varepsilon_{i1}, \ldots, \varepsilon_{iM})'$. In compact notation, the unobserved components of the model are denoted $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_N)'$ and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_N)'$, respectively.

In classical EFA, the dimension of the factor covariance matrix is estimated using a variety of criteria. Various ad hoc rules for allocating meaurements to factors are used (Gorsuch, 2003). As in classical EFA we assume that the measurements are *dedicated*, i.e., that each measurement loads on at most a single factor. If a measurement does not load on any factor the measurement is discarded from the model. In classical EFA, measurements that load on multiple factors are also discarded. Our analysis improves on this procedure. The position of the non-zero elements in the factor loading matrix is not fixed *a priori*, but is determined during estimation along with the number of factors, which is not imposed but estimated. In addition, we use all measurements.

To indicate how measurements are uniquely allocated to the factors in $\boldsymbol{\theta}$, we use a matrix of binary indicators $\boldsymbol{\Delta}$ with the same dimensions as the factor loading matrix $\boldsymbol{\alpha}$. Each row of $\boldsymbol{\Delta}$ indicates on which latent factor the corresponding measurement loads. For example, if the m^{th} measurement is associated with factor k, then the m^{th} row Δ_m is the indicator vector e_k:

$$\Delta_m = (0, \ldots, 0, \underbrace{1}_{\text{k}^{\text{th}}\text{ element}}, 0, \ldots, 0) \equiv e_k. \tag{2}$$

When a measurement does not load on any factor, the corresponding row of $\boldsymbol{\Delta}$ only contains zeros (denoted vector e_0). Under our assumptions, no measurements may load on more than one factor, though any measurement may load on no factors, i.e., $\sum_k \Delta_{mk} \leq 1$.

Since neither the number of factors nor the structure of the factor loading matrix are specified *a priori*, the indicator matrix $\boldsymbol{\Delta}$ is one of the unknowns of the model to be estimated

[1] We only consider continuous and binary measurements in this paper, because of our empirical application where such measurements are available. The methodology can be extended to any other types of discrete measurements with an underlying continuous latent variable.

from the data. This matrix representation is convenient for the implementation of the factor search procedure introduced in Section 3. The values assumed by $\boldsymbol{\Delta}$ determine how measurements are allocated to the different dedicated factors, which factors are shut down (zero columns of $\boldsymbol{\Delta}$), and the number of factors underlying the data (the number of non-zero columns). Indicator matrix $\boldsymbol{\Delta}$ has been widely used in variable selection models (Geweke, 1996; George and McCulloch, 1997). In the framework of factor analysis, it is used by Carvalho et al. (2008); Frühwirth-Schnatter and Lopes (2012); Chen et al. (2011). Our approach departs from these papers because we use a dedicated structure for the factor loading matrix and correlated factors.

2.2 Classical Identification

This section presents and discusses classical identification strategies used in factor analysis. We introduce a theorem for the identification of dedicated factor models of varying dimensions, explain how to apply classical identification criteria to Bayesian inference and outline the benefits of this approach.

General Identification Strategy. We center the unobserved components of the model, θ_i and ε_i, at:

$$\begin{aligned} E(\theta_i) &= 0, & \text{Cov}(\theta_i) &= \boldsymbol{\Omega}, & & \\ E(\varepsilon_i) &= 0, & \text{Cov}(\varepsilon_i) &= \boldsymbol{\Sigma}, & \boldsymbol{\Sigma} &= \text{diag}(\sigma_1^2, \ldots, \sigma_M^2). \end{aligned} \tag{3}$$

The components of ε_i are mutually uncorrelated. Conditional on X_i, the latent factors are the only source of correlation among the measurements.[2] The latent factors are assumed to be independent of the error terms and of the covariates, i.e., $\theta_i \perp\!\!\!\perp \varepsilon_i$ and $\theta_i \perp\!\!\!\perp X_i$. In addition, we assume that for all measurements, the variances of the idiosyncratic errors are positive, i.e., $\sigma_m^2 > 0$. In the equations corresponding to the latent variables generating the binary measurements, these variances are set to 1, i.e., $\sigma_m^2 = 1$. Without further information, the scales of the corresponding latent variables $Y_i^\star$ are not identified.

We follow traditions in factor analysis and only consider identification based on population means and covariance matrices.[3] Our assumptions imply the following covariance

[2] Cunha and Heckman (2008), Appendix A, show how the measurements can be interpreted as derived demand functions for producing factors θ.

[3] Bonhomme and Robin (2010) consider identification of factor models based on higher order moments.

structure for the latent variables $Y_i^\star$:

$$\text{Cov}(Y_i^\star \mid X_i) = \boldsymbol{\alpha}\boldsymbol{\Omega}\boldsymbol{\alpha}' + \boldsymbol{\Sigma}, \tag{4}$$

where the diagonal elements of $\boldsymbol{\Sigma}$ corresponding to the latent variables underlying discrete measurements are restricted to be 1. Identification of the parameters $\boldsymbol{\alpha}$,$\boldsymbol{\Omega}$, and $\boldsymbol{\Sigma}$ from $\text{Cov}(Y_i^\star \mid X_i)$ requires further restrictions.

To secure classical identification, conditions are required that guarantee the existence of a unique solution for the idiosyncratic variances $\boldsymbol{\Sigma}$ (the *uniqueness* problem). This problem is sometimes addressed by verifying that the number of latent factors does not exceed the Ledermann bound, i.e., $K \leq \phi(M) = (2M + 1 - \sqrt{8M+1})/2$ (Ledermann, 1937; Bekker and ten Berge, 1997).[4] Anderson and Rubin (1956, Theorem 5.6) establish that at least three non-zero elements are required in each column of the factor loading matrix to achieve uniqueness.

Given identifiability of $\boldsymbol{\Sigma}$, further conditions are needed to guarantee the existence of a unique solution for the factor loading matrix $\boldsymbol{\alpha}$ and the covariance matrix of the factors $\boldsymbol{\Omega}$. The "rotation problem" stems from the fact that the covariance in equation (4) remains unchanged after assigning $\widetilde{\boldsymbol{\alpha}} = \boldsymbol{\alpha}\boldsymbol{P}$ and $\widetilde{\theta}_i = \boldsymbol{P}^{-1}\theta_i$, for any arbitrary nonsingular matrix $\boldsymbol{P}$ of dimension $(K \times K)$.

To solve this problem, various restrictions and normalizations are used in the literature. First of all, it is necessary to deal with the scaling issue. In the framework of our dedicated model, we assume that the covariance matrix of the factors, $\boldsymbol{\Omega}$, is of full rank. We fix the variances of the factors to 1 to set the scales of the loadings:

$$\text{rank}(\boldsymbol{\Omega}) = K, \qquad\qquad \text{diag}(\boldsymbol{\Omega}) = \iota_K, \tag{5}$$

where $\iota_K = (1, \ldots, 1)'$ is a vector of ones of length K. We denote by $\boldsymbol{R}$ the correlation matrix of the factors to distinguish it from the covariance matrix $\boldsymbol{\Omega}$. These restrictions leave the factor loading matrix completely free, compared to alternative and more conventional identifying restrictions that fix one loading to 1 in each column of $\boldsymbol{\alpha}$ (e.g., Carneiro et al., 2003; Heckman et al., 2006). Such identifying strategies, however, cannot be implemented with our algorithm, as we do not know *a priori* the number of factors, nor how the measurements are allocated to the factors. As a consequence, it is impossible to fix any loadings *a priori*.

[4]The Ledermann bound simply requires that the number of equations be greater than or equal to the number of model unknowns.

Additional assumptions are required to identify the model and rule out remaining rotation problems. Anderson and Rubin (1956) postulate, among other specifications, lower triangularity for the upper square submatrix of $\boldsymbol{\alpha}$, and versions of this specification have been widely used in econometrics (see, e.g., Geweke and Zhou, 1996; Aguilar and West, 2000; Carneiro et al., 2003). In the context of sparse factor modeling, however, the configuration of the zero elements in the factor loading matrix plays a crucial role for the identifiability of the model, as a minimum number of non-zero loadings is required in each column of the factor loading matrix (Anderson and Rubin, 1956). As a consequence, imposing zero elements only on the upper triangular part of $\boldsymbol{\alpha}$ may not be sufficient for identification. Given that any loading in the lower triangular part of the matrix may become equal to zero, too many zeros may jeopardize identification. Most applications in sparse factor modeling rely on a lower triangular structure of $\boldsymbol{\alpha}$ and do not address these potential identifiability issues (West, 2003; Lopes and West, 2004; Lucas et al., 2006; Carvalho et al., 2008).[5] Exceptions are Carneiro et al. (2003) and Frühwirth-Schnatter and Lopes (2012), who use classical identification criteria as an integral part of Bayesian inference schemes.

The present paper addresses these problems and achieves more flexibility in this respect. At the same time, it solves both the uniqueness and rotation problems, apart from trivial rotations to be discussed below. We assume a dedicated structure and that factors are either loaded on *at least three measurements* or not loaded on any measurements, in which case they are discarded from the model.[6] Since measurements appear in blocks of dedicated measurements, it is unlikely that the first K measurements are actually dedicated to the K different factors, as suggested by a lower triangular loading matrix with non-zero entries on the main diagonal.

In the framework of a dimension-varying model where the structure of the factor loading matrix in terms of zero elements is not known a priori, more general identification conditions are required and are now presented.

Identification of a Dimension-Varying Model. The following theorem introduces sufficient conditions for identifiability of a dedicated factor model when the allocation of measurements to factors is unknown.

[5]Most of these papers deal with high-dimensional factor models, where factors are usually loaded by a myriad of measurements. In such cases, these identification problems are not a concern in practice. However, in smaller models where these problems may arise, it is important to address them appropriately.

[6]In our framework with correlated factors, only two measurements are required for each factor, as long as the correlation between the corresponding factors and the other factors is not zero (Cunha et al., 2010). We do not use these conditions though, because we allow for zero correlations across factors.

Theorem 1. *Consider a dedicated factor model with K factors satisfying condition (5). Furthermore, assume that the number of non-zero elements in the k^{th} column of $\boldsymbol{\Delta}$, $n_k(\boldsymbol{\Delta}) = \sum_{m=1}^{M} \Delta_{mk}$, is either equal to 0 or at least equal to 3 for all $k = 1, \ldots, K$:*

$$n_k(\boldsymbol{\Delta}) \geq 3 \quad or \quad n_k(\boldsymbol{\Delta}) = 0, \qquad \forall k = 1, \ldots, K. \tag{6}$$

Then the factor model is identified up to trivial rotations. More specifically, the indicator matrix $\boldsymbol{\Delta}$ is identifiable up to an arbitrary permutation of the columns, whereas the factor loading matrix $\boldsymbol{\alpha}$ and the submatrix of the correlation matrix $\boldsymbol{\Omega}$ corresponding to the non-zero columns of $\boldsymbol{\Delta}$ are identifiable up to the same permutation of the columns and up to a sign switch for each column.

Proof. First, we prove identifiability of $\boldsymbol{\Sigma}$. Anderson and Rubin (1956, Theorem 5.1) present a sufficient condition for identifiability of $\boldsymbol{\Sigma}$: the "row deletion" property which states that if any row of $\boldsymbol{\alpha}$ is deleted, there remain two disjoint matrices that are of the same rank as $\boldsymbol{\alpha}$. For a dedicated factor model, rank($\boldsymbol{\alpha}$) is equal to the number K_1 of non-zero columns of $\boldsymbol{\alpha}$. It is easy to verify that condition (6) implies the row deletion property, because regardless of whether a zero or a non-zero row is deleted, it is guaranteed that in each of the K_1 non-zero columns at least two non-zero factor loadings are still present. Hence, both the diagonal sub-matrix $\boldsymbol{\alpha}_1$ constructed from the top non-zero elements in each non-zero column as well as the remaining sub-matrix $\boldsymbol{\alpha}_2$ still has K_1 non-zero columns, and consequently the rank is equal to K_1.

Next, consider any alternative representation $(\boldsymbol{\alpha}^\star, \boldsymbol{\Omega}^\star, \boldsymbol{\Sigma}^\star)$ of $\mathrm{Cov}(Y_i^\star \mid X_i)$, defined in (4), where $\boldsymbol{\alpha}^\star$ and $\boldsymbol{\Omega}^\star$ obey conditions (5) and (6). Identifiability of $\boldsymbol{\Sigma}$ implies $\boldsymbol{\Sigma}^\star = \boldsymbol{\Sigma}$, hence, identifiability of $\boldsymbol{\alpha\Omega\alpha}'$, i.e.:

$$\boldsymbol{\alpha\Omega\alpha}' = \boldsymbol{\alpha}^\star \boldsymbol{\Omega}^\star (\boldsymbol{\alpha}^\star)'. \tag{7}$$

Due to the dedicated nature of the factor loading matrix, in both representations at most one element α_m and $\alpha_m^\star$ is different from 0 in each row m. From the restrictions on the diagonal elements of the covariance matrix of the factors, we obtain the following relationship between α_m and $\alpha_m^\star$ from the diagonal elements of the covariance matrices appearing in (7):

$$\alpha_m^2 = (\alpha_m^\star)^2, \qquad m = 1, \ldots, M. \tag{8}$$

Thus α_m is zero if and only if $\alpha_m^\star$ is equal to zero. Hence, the subset of measurements that do not load on any factors is the same for both solutions. Thus, further investigations may be limited to dedicated measurements, where both α_m and $\alpha_m^\star$ are different from 0. It follows

immediately from equation (8) that the factor loadings of any dedicated measurement are the same for both solutions, apart from sign switching. However, this does not necessarily imply that the measurement is dedicated to the same factor, i.e., δ_m might be different from $\delta_m^\star$, where δ_m and $\delta_m^\star$ indicate the position of the non-zero elements of the indicator vectors Δ_m and $\Delta_m^\star$, respectively.

For further investigation, consider the off-diagonal elements of the covariance matrices appearing in (7), defining the covariance between any pair (m, l) of dedicated measurements:

$$\alpha_m \Omega_{\delta_m,\delta_l} \alpha_l = \alpha_m^\star \Omega^\star_{\delta_m^\star,\delta_l^\star} \alpha_l^\star. \tag{9}$$

It follows immediately from (8) and (9) that

$$\Omega^2_{\delta_m,\delta_l} = (\Omega^\star_{\delta_m^\star,\delta_l^\star})^2. \tag{10}$$

Now consider any pair (m, l) of measurements that are dedicated to the same factor j in the representation corresponding to $\boldsymbol{\alpha}$, i.e., $\delta_m = \delta_l = j$, and $\Omega_{\delta_m,\delta_l} = \Omega_{jj} = 1$ because of the restriction defined in equation (5). Assume that these measurements are *not* dedicated to the same factor in the representation corresponding to $\boldsymbol{\alpha}^\star$, i.e., $\delta_m^\star \neq \delta_l^\star$. Equation (10) implies that $\Omega^\star_{\delta_m^\star,\delta_l^\star} = \pm|\Omega_{jj}| = \pm 1$, and as a consequence the two factors corresponding to the columns $\delta_m^\star$ and $\delta_l^\star$ of $\boldsymbol{\Delta}$ have to be perfectly correlated in the alternative representation, which contradicts the full rank condition for $\boldsymbol{\Omega}^\star$ given by equation (5). Hence, it follows that $\delta_m^\star = \delta_l^\star$ whenever $\delta_m = \delta_l$, meaning that the same subset of measurements is dedicated to a particular factor in both representations.

This implies that $(\boldsymbol{\Delta}, \boldsymbol{\alpha})$ and $(\boldsymbol{\Delta}^\star, \boldsymbol{\alpha}^\star)$ have the same number K_1 of non-zero columns. However, the position of the non-zero columns is not unique and $\boldsymbol{\Delta}$ is identifiable up to column switching, i.e.:

$$\boldsymbol{\Delta}^\star = \boldsymbol{\Delta} \boldsymbol{P}_\rho, \tag{11}$$

where the (orthonormal) rotation matrix $\boldsymbol{P}_\rho$ corresponds to a permutation matrix of the columns. Furthermore, $\boldsymbol{\alpha}$ is identified up to the same permutation of the columns as well as a possible sign switching, see (8):

$$\boldsymbol{\alpha}^\star = \boldsymbol{\alpha} \boldsymbol{P}_\rho \boldsymbol{P}_\pm, \tag{12}$$

where $\boldsymbol{P}_\pm = \mathrm{diag}(\pm 1, \ldots, \pm 1)$.

Finally, let $\boldsymbol{\Omega}_1$ and $\boldsymbol{\alpha}_1$ be, respectively, the submatrix of the correlation matrix $\boldsymbol{\Omega}$ and the factor loading matrix $\boldsymbol{\alpha}$ corresponding to the non-zero columns of $\boldsymbol{\Delta}$. From (7) and (12) it follow that[7]

$$\boldsymbol{\alpha}_1\boldsymbol{\Omega}_1\boldsymbol{\alpha}_1' = \boldsymbol{\alpha}_1^\star\boldsymbol{\Omega}_1^\star(\boldsymbol{\alpha}_1^\star)' = \boldsymbol{\alpha}_1(\boldsymbol{P}_\rho)_1(\boldsymbol{P}_\pm)_1\boldsymbol{\Omega}_1^\star(\boldsymbol{P}_\pm)_1'(\boldsymbol{P}_\rho)_1'\boldsymbol{\alpha}_1',$$

and, hence:

$$\boldsymbol{\Omega}_1^\star = (\boldsymbol{P}_\pm)_1'(\boldsymbol{P}_\rho)_1'\boldsymbol{\Omega}_1(\boldsymbol{P}_\rho)_1(\boldsymbol{P}_\pm)_1.$$

This implies identifiability of $\boldsymbol{\Omega}_1$ up to column switching and sign switching. □

Theorem 1 only achieves identification of the submatrix of $\boldsymbol{\Omega}$ corresponding to the non-zero columns of $\boldsymbol{\Delta}$. Indeed, the covariances between the unidentified factors—those that are not loaded by any factors—as well as the covariances between the unidentified factors and the dedicated factors, are not identifiable in the overall model. However, only the latent factors actually underlying the measurements are of interest, so that this lack of identification is not a concern.

Application of Classical Identification Criteria to Bayesian Inference. Identifiability condition (6) is easy to check and very convenient from a computational point of view, as it only applies to the indicator matrix $\boldsymbol{\Delta}$, and is therefore easily incorporated in the algorithm introduced in the next section. To do so, we design a prior distribution for $\boldsymbol{\Delta}$ that restricts the sampler to explore regions of the parameter space corresponding to identified models only (i.e., only indicator matrices satisfying the identification conditions are sampled).

No further restrictions need to be enforced *a priori* to resolve the remaining trivial rotation problems, outlined in the proof of Theorem 1 in equations (11) and (12), namely identifiability up to *sign switching* and *column switching*. The former appears when the signs of the factor loadings in a given column of $\boldsymbol{\alpha}$ and the sign of the corresponding factor θ_i are switched simultaneously. The latter arises from the fact that there is no natural ordering of the columns of $\boldsymbol{\alpha}$—they can be permuted, along with the corresponding latent factors θ_i, without altering the covariance structure of the measurements. These two trivial identifiability problems, however, can be addressed *a posteriori* by reordering the columns of the loadings matrix and switching the signs of the loadings appropriately (see Subsection 3.4).

[7]Similarly, $(\boldsymbol{P}_\pm)_1$ and $(\boldsymbol{P}_\rho)_1$ are, respectively, the submatrices of $\boldsymbol{P}_\pm$ and $\boldsymbol{P}_\rho$ corresponding to the non-zero columns of $\boldsymbol{\Delta}$.

Concerning the maximum number of factors, equation (6) implies the following upper bound on the number of factors that can be extracted from M measurements:[8]

$$K \leq K^{\max} = \min\left\{\frac{M}{3},\ \phi(M)\right\}.$$

Hence, for a dedicated factor model with $M \geq 4$ the requirement of at least three measurements loading on each dedicated factor becomes stronger than the Ledermann bound.[9]

For the Bayesian inference pursued in this paper, a complete distributional specification of model equation (1) is required, which goes beyond specifying first- and second order moments of θ_i and ε_i as in equation (3). Cunha et al. (2010) establish nonparametric identifiability of the distributions of $\boldsymbol{\theta}$ and $\boldsymbol{\varepsilon}$. To adapt their results to our model, we would have to use a Bayesian nonparametric approach (Ghahramani et al., 2007; Paisley and Carin, 2009; Bhattacharya and Dunson, 2011). To avoid the substantial computational challenge associated with such a Bayesian nonparametric approach, we invoke the following normality assumptions on the latent factors and on the error terms:

$$\theta_i \sim \mathcal{N}(0;\ \boldsymbol{R}), \qquad \varepsilon_i \sim \mathcal{N}(0;\ \boldsymbol{\Sigma}),$$

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

Practical Bayesian inference would not necessarily impose the strict identifying restrictions presented in this section, as they are not required to conduct inference. Learning about model parameters can indeed take place, even if the model is not identified in a classical sense (Poirier, 1998). However, a lack of identification can impair interpretation, if for instance spurious factors are generated. This contradicts the goal of Bayesian exploratory factor analysis that seeks to uncover a structure of the model that can be easily interpreted. Nevertheless, this goal can be restored by constraining the sampler to stay in regions of the parameter space where only (classically) identified models are generated. The next section introduces our prior specification, and explains how these classical identification conditions are integrated into our Bayesian inference procedure.

2.3 Elements of Prior Specification

2.3.1 The Prior on the Indicators

The allocation of the measurements to groups of dedicated measurements can be interpreted as a mixture problem with unknown, but finite, number of components. Let τ_k denote the

[8] This upper bound is not strictly required in a panel context, see Cunha et al. (2010).
[9] See Frühwirth-Schnatter and Lopes (2012).

probability that a measurement loads on factor k. It does not load on any factor if $k = 0$. For each row Δ_m of $\boldsymbol{\Delta}$,[10] for $m = 1, \ldots, M$, we assume:

$$\Pr(\Delta_m = e_k \mid \tau_k) = \tau_k, \qquad k = 0, 1, \ldots, K, \tag{13}$$

where e_k is the indicator vector of length K as defined in equation (2), and $\sum_{k=0}^{K} \tau_k = 1$.

The allocation of each measurement to one of the dedicated groups of measurements can be seen as a two-step decision, in which we incorporate a *hierarchical prior* on the indicators $\boldsymbol{\Delta}$. First, with probability τ_0 we assume that a measurement does not load on any factor. In this case, it is uncorrelated with the other measurements and does not contribute to the extraction of the factors. It is thus implicitly discarded from the model. In the opposite case, this measurement loads on a latent factor with probability $1 - \tau_0$. Conditional on this event, it is then allocated to one of the K groups of dedicated measurements according to a set of probabilities $\tau^\star = (\tau_1^\star, \ldots, \tau_K^\star)'$, with $\sum_{k=1}^{K} \tau_k^\star = 1$. The probabilities of the different events can thus be written as

$$\tau = (\tau_0, \tau_1, \ldots, \tau_K)' = (\tau_0, (1 - \tau_0)\tau_1^\star, \ldots, (1 - \tau_0)\tau_K^\star)'. \tag{14}$$

To conduct Bayesian inference, we have to place prior distributions on these parameters. We assume the following:

$$\tau_0 \sim \mathcal{B}eta(\kappa_0; \xi_0)\,, \qquad \tau^\star = (\tau_1^\star, \ldots, \tau_K^\star)' \sim \mathcal{D}ir(\kappa_1, \ldots, \kappa_K)\,, \tag{15}$$

where the Beta distribution for τ_0 is defined on the support $[0; 1]$ and has mean $\kappa_0/(\kappa_0 + \xi_0)$. It can be specified so as to obtain more or less mass toward 0 or 1, depending on our prior knowledge about the number of measurements that should be discarded from the analysis. The Dirichlet distribution on the weights $\tau^\star$ is quite standard in mixture modeling (see e.g. Frühwirth-Schnatter, 2006).

Unfortunately, the indicator probabilities specified in equation (13), equipped with the prior distributions defined in equation (15), result in a prior distribution $p(\boldsymbol{\Delta}) = \int p(\boldsymbol{\Delta} \mid \tau)p(\tau)d\tau$ that does *not* guarantee identification of the model. To secure identification, as discussed in Subsection 2.2, the prior needs to incorporate the restriction that at least three dedicated measurements have to load on each latent factor. This can be achieved by restricting the distribution of $\boldsymbol{\Delta}$ to the subset $\mathcal{D}$ of matrices that correspond to an identified

[10] $\boldsymbol{\Delta}$ is the matrix of binary indicators with the same dimensions as the factor loading matrix $\boldsymbol{\alpha}$.

model:[11]

$$p(\boldsymbol{\Delta} \mid \tau, \mathcal{D}) \propto \left(\prod_{k=0}^{K} \tau_k^{n_k(\boldsymbol{\Delta})} \right) \delta_{\mathcal{D}}(\boldsymbol{\Delta}), \tag{16}$$

where $n_k(\boldsymbol{\Delta}) = \sum_{m=1}^{M} \Delta_{mk}$ is the number of elements in the set of measurements dedicated to factor k, for $k = 1, \ldots, K$, $n_0(\boldsymbol{\Delta})$ is the number of measurements that do not load on any factors, and $\delta_{\mathcal{D}}(\boldsymbol{\Delta})$ is the Dirac measure that is equal to 1 if $\boldsymbol{\Delta}$ belongs to $\mathcal{D}$, to 0 otherwise. The subset of indicator matrices $\mathcal{D}$ can be formally expressed as:

$$\mathcal{D} = \left\{ \quad \boldsymbol{\Delta} \quad \middle| \quad \sum_{k=1}^{K} \Delta_{mk} \leq 1 \quad \forall\, m = 1, \ldots, M, \quad n_k(\boldsymbol{\Delta}) \geq 3 \text{ or } = 0 \quad \forall\, k = 1, \ldots, K \right\}.$$

More flexible hierarchical prior specification. As an alternative, it is possible to specify individual parameters τ_{0m} for the measurements, to make the probability of inclusion into the model measurement-specific and independent of the other measurements. The remaining indicator probabilities $\tau^\star$ are specified to be common to all measurements as before, i.e., $\tau^\star \sim \mathcal{D}ir(\kappa_1, \ldots, \kappa_K)$.[12] This minor modification implies that for each measurement $m = 1, \ldots, M$, we specify:

$$\tau_m = (\tau_{0m}, (1 - \tau_{0m})\tau_1^\star, \ldots, (1 - \tau_{0m})\tau_K^\star)',$$

and assume that $\tau_{0m} \sim \mathcal{B}eta(\kappa_0;\, \xi_0)$.

Our Monte Carlo studies show that this simple modification of the prior considerably improves the ability of our algorithm to find the measurements that do not load on any factors (see Subsection 4.1). This result also becomes clear when we derive our MCMC sampler. When the same τ_0 is specified across measurements, its posterior distribution decreases with the number of correlated measurements. This makes it difficult to retrieve the number of uncorrelated measurements, as their posterior probability is forced to be the same for all measurements and can become very small in large models.[13]

[11] The normalizing constant of this distribution can be derived in closed-form solution, but is not required in our analysis.

[12] The parameters $\tau^\star$ could also be specified as measurement-specific, but our tests indicated that this specification led to model overfitting.

[13] More precisely, the posterior mean of τ_0 decreases if the number of measurements M increases while the number of uncorrelated measurements remains fixed, see equation (A6).

2.3.2 The Prior on the Idiosyncratic Variances

For all continuous measurements Y_{im}, we specify an inverse-Gamma prior distribution on the variances of the idiosyncratic error terms:

$$\sigma_m^2 \sim \mathcal{G}^{-1}\big(c_0;\, C_m^0\big)\,, \qquad\qquad m \in \mathcal{I}_{\text{cont}},$$

where $\mathcal{I}_{\text{cont}} \subset (1, \ldots, M)$ is the set of indices corresponding to the continuous measurements, and c_0 and C_m^0 are scalar parameters denoting the shape and the scale of the distribution. The inverse-Gamma distribution is defined on the positive support and therefore guarantees that the variances cannot be negative, preventing some idiosyncratic variances from lying outside of the admissible parameter range, a phenomenon known as a Heywood case (after Heywood, 1931), in the likelihood analysis of factor models. To specify the hyperparameters, we follow Frühwirth-Schnatter and Lopes (2012) who develop a data-driven prior that makes use of the observed covariance matrix $\boldsymbol{S}_{Y_{\text{cont}}}$ of the measurements and specify the scale parameter such that:

$$\sigma_m^2 \sim \mathcal{G}^{-1}\left(c_0;\, \frac{c_0 - 1}{\big(\boldsymbol{S}_{Y_{\text{cont}}}^{-1}\big)_{mm}}\right), \tag{17}$$

where $\big(\boldsymbol{S}_{Y_{\text{cont}}}^{-1}\big)_{mm}$ is the m^{th} diagonal element of the inverse of the empirical covariance matrix of the continuous measurements Y_{cont}.[14]

2.3.3 The Prior on the Factor Loadings

The indicator matrix $\boldsymbol{\Delta}$ determines the factors to which the different measurements are dedicated. A direct consequence is that a given factor loading α_{mk}, in row m and column k of $\boldsymbol{\alpha}$, will either be equal to zero (if $\Delta_{mk} = 0$), or follow a prior distribution that needs to be specified (if $\Delta_{mk} = 1$). Following the usual assumptions in Bayesian factor analysis, we assume that the factor loadings are independent across measurements and adopt the usual normal-inverse-Gamma family as prior distribution, meaning that conditional on knowing σ_m^2 and Δ_m, the only non-zero factor loading $\alpha_m^{\boldsymbol{\Delta}}$ in the m^{th} row of the factor loading matrix $\boldsymbol{\alpha}$ is conditionally normal:

$$\alpha_m^{\boldsymbol{\Delta}} \mid \sigma_m^2 \sim \mathcal{N}\big(a_m^0;\, A_m^0 \sigma_m^2\big)\,, \tag{18}$$

[14] Note that if $N < M_{\text{cont}}$, where M_{cont} is the number of continuous measurements, the empirical covariance matrix is not positive-definite and therefore this approach cannot be applied. A prior distribution with pre-specified scale parameter C_m^0 has to be used in this case. See Frühwirth-Schnatter and Lopes (2012) for details.

where a_m^0 and A_m^0 are scalar parameters denoting the prior mean and the scale of the variance, respectively.

The normal-inverse Gamma family has several advantages in the present context. First, it allows us to integrate the joint posterior distribution $p(\boldsymbol{\Delta}, \boldsymbol{\alpha}, \boldsymbol{\Sigma} \mid \boldsymbol{Y}^\star, \boldsymbol{\theta}, \boldsymbol{\beta}, \tau)$ over $\boldsymbol{\alpha}$ and $\boldsymbol{\Sigma}$, making sampling from $p(\boldsymbol{\Delta} \mid \boldsymbol{Y}^\star, \boldsymbol{\theta}, \boldsymbol{\beta}, \tau)$ possible, see Subsection 3.1.1. Second, the prior defined in (18) induces a more diffuse prior on the factor loadings when measurement error is larger and implies the following prior distribution for the amount of variance explained by the corresponding dedicated factor,

$$\frac{(\alpha_m^{\boldsymbol{\Delta}})^2}{(\alpha_m^{\boldsymbol{\Delta}})^2 + \sigma_m^2} = \frac{(\tilde{\alpha}_m^{\boldsymbol{\Delta}})^2}{(\tilde{\alpha}_m^{\boldsymbol{\Delta}})^2 + 1},$$

where $\tilde{\alpha}_m^{\boldsymbol{\Delta}} \sim \mathcal{N}(a_m^0;\, A_m^0)$. This ratio has the same prior distribution for any two dedicated measurements m and l, where $a_m^0 = a_l^0$ and $A_m^0 = A_l^0$.

By integrating out the indicators, the marginal prior distribution of α_{mk} turns out to be a mixture of a point mass at zero and a normal distribution with a fixed-scale variance. Such prior distributions have previously been used in the framework of sparse factor modeling, as they allow model shrinkage (West, 2003; Lucas et al., 2006; Carvalho et al., 2008; Frühwirth-Schnatter and Lopes, 2012). The exact form of the mixture is more difficult to derive analytically in our case, because of the identifying restrictions on $\boldsymbol{\Delta}$. Nevertheless, we only need the conditional prior distribution specified in equation (18) for Bayesian inference, as only the non-zero factor loadings need to be sampled.

2.3.4 The Prior on the Regression Coefficients

Let $\boldsymbol{\beta} = (\beta_1 \dots \beta_M)'$, where β_m' corresponds to the m^{th} row of the matrix of regression coefficients $\boldsymbol{\beta}$. Each of these vectors is assumed to be *a priori* normally distributed:

$$\beta_m \sim \mathcal{N}\left(b_m^0;\, \boldsymbol{B}_m^0\right), \qquad\qquad m = 1, \dots, M,$$

where b_m^0 is a vector of prior mean parameters of length Q, and $\boldsymbol{B}_m^0$ is a $(Q \times Q)$–dimensional prior covariance matrix.

2.3.5 The Prior on the Correlation Matrix of the Factors

The correlation matrix of the factors is sampled through marginal data augmentation. Before turning to the details of this procedure in Subsection 3.2.1, it is important to understand how

the distribution of the covariance matrix $\boldsymbol{\Omega}$ of the latent factors is related to the distribution of their variances and to the distribution of the corresponding correlation matrix $\boldsymbol{R}$.

Given the decomposition $\boldsymbol{\Omega} = \boldsymbol{\Lambda}^{\frac{1}{2}} \boldsymbol{R} \boldsymbol{\Lambda}^{\frac{1}{2}}$, where $\boldsymbol{\Lambda} = \text{diag}(\Lambda_1, \ldots, \Lambda_K)$ contains the variances of the factors, Zhang et al. (2006) show that if it is assumed that $\boldsymbol{\Omega} \sim \mathcal{W}_K^{-1}(\nu; \boldsymbol{S})$, an inverse-Wishart distribution with ν degrees of freedom, where $\nu - K + 1 > 0$, and scale matrix $\boldsymbol{S}$, the joint distribution of $\boldsymbol{\Lambda}$ and $\boldsymbol{R}$ can be obtained through the transformation from $\boldsymbol{\Omega}$ to $(\boldsymbol{\Lambda}, \boldsymbol{R})$ using the corresponding Jacobian $\mathcal{J}_{(\boldsymbol{\Omega} \to \boldsymbol{\Lambda}, \boldsymbol{R})} = |\boldsymbol{\Lambda}|^{\frac{K-1}{2}}$:[15]

$$\begin{aligned} p(\boldsymbol{\Lambda}, \boldsymbol{R} | \boldsymbol{S}) &= \mathcal{J}_{(\boldsymbol{\Omega} \to \boldsymbol{\Lambda}, \boldsymbol{R})}\, p(\boldsymbol{\Omega}), \\ &= c\, |\boldsymbol{S}|^{\frac{\nu}{2}} |\boldsymbol{\Lambda}|^{-\frac{\nu}{2} - 1} |\boldsymbol{R}|^{-\frac{(\nu+K+1)}{2}} \exp\left\{ -\frac{1}{2} \text{tr}\left(\boldsymbol{S} \boldsymbol{\Lambda}^{-1} \boldsymbol{R}^{-1} \right) \right\}. \end{aligned} \tag{19}$$

The hyper parameter $\boldsymbol{S}$ in the inverted-Wishart prior chosen for $\boldsymbol{\Omega}$ can either be assumed to be fixed or a hyper prior $p(\boldsymbol{S})$ may be assumed for $\boldsymbol{S}$. Following Huang and Wand (2013), $\boldsymbol{S} = \text{diag}(s_1, \ldots, s_K)$ is assumed to be a nonsingular diagonal matrix where the individual variances follow a Gamma distribution,

$$s_k \sim \mathcal{G}\left(\frac{1}{2}; \frac{1}{2\, \nu^\star A_k^2} \right), \qquad \text{for } k = 1, \ldots, K, \tag{20}$$

with $\nu^\star = \nu - K + 1$.[16] For the special case where the scale matrix $\boldsymbol{S} = \text{diag}(s_1, \ldots, s_K)$ is a nonsingular diagonal matrix, being either fixed or random, the marginal distribution of $\boldsymbol{R}$ can be derived in closed-form solution by integrating out $\boldsymbol{\Lambda}$ of equation (19) (Zhang et al., 2006, see also Barnard et al., 2000, Section 2.2):

$$p(\boldsymbol{R} | \boldsymbol{S}) = \int p(\boldsymbol{\Lambda}, \boldsymbol{R} | \boldsymbol{S}) d\boldsymbol{\Lambda} = 2^{\nu K/2} \Gamma_K(\nu/2) |\boldsymbol{R}|^{-\frac{(\nu+K+1)}{2}} \left(\prod_k r^{kk} \right)^{-\frac{\nu}{2}} = p(\boldsymbol{R}), \tag{21}$$

where r^{kk} is the k^{th} diagonal element of the inverse of $\boldsymbol{R}$.

It should be noted that the marginal density $p(\boldsymbol{R})$ of the correlation matrix $\boldsymbol{R}$ given by (21) does not depend on $\boldsymbol{S}$, leaving the degrees of freedom parameter ν as the only hyper-

[15] The inverse-Wishart distribution is parameterized as follows:

$$p(\boldsymbol{\Omega}) = c\, |\boldsymbol{S}|^{\frac{\nu}{2}} |\boldsymbol{\Omega}|^{-\frac{(\nu+K+1)}{2}} \exp\left\{ -\frac{1}{2} \text{tr}\left(\boldsymbol{S} \boldsymbol{\Omega}^{-1} \right) \right\},$$

with normalizing constant

$$c = 1/(2^{\nu K/2} \Gamma_K(\nu/2)),$$

where $\Gamma_K(\cdot)$ is the generalized Gamma function.

[16] The Gamma distribution in equation (20) is parameterized such that the expectation of s_k is equal to $\nu^\star A_k^2$.

parameter of this prior. Barnard et al. (2000) discuss how to specify the hyper-parameter ν, and show that taking $\nu = K + 1$ (i.e., $\nu^\star = 2$) results in a uniform marginal distribution of the individual correlations. Increasing the hyper-parameter ν induces bell-shaped distributions by assigning a prior probability to neighborhoods of ± 1 that goes to 0 as ν increases, bounding the correlations away from ± 1.

The degrees of freedom ν of the inverse-Wishart distribution plays an important role in the tuning of our algorithm. Intuitively, the stronger the correlation among the latent factors *a priori*, the more likely a larger number of latent factors will be favored. Some factors might indeed be split into several highly-correlated factors when the prior allows for high correlations. This "factor splitting" problem is at odds with our goal of generating a sparse and interpretable structure, as it can result in an overfitting of the number of factors, where some of them appear to be redundant in explaining the data.

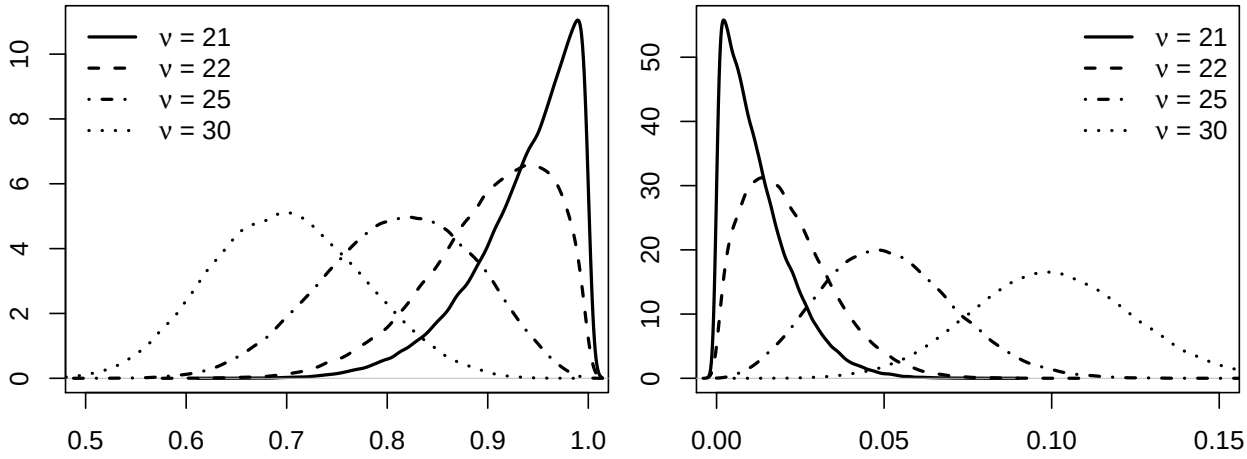
In addition, according to Theorem 1, the full rank condition for the correlation matrix $\boldsymbol{R}$ also plays an important role for the identification of the indicator matrix $\boldsymbol{\Delta}$. If only a few measurements load on a particular factor, then the information contained in the measurements might not be sufficient to bound the posterior distribution away from regions where $\boldsymbol{R}$ is rank deficient. The prior on $\boldsymbol{R}$ secures the identifiability of $\boldsymbol{\Delta}$.

To address these issues, the degrees of freedom ν of the prior on the correlation matrix can be tuned to bound the posterior away from regions of unidentifiability. For further illustration, Figure 1 shows the marginal prior distribution $p(\max_{j \neq k} |R_{jk}|)$ of the largest correlation coefficient appearing in $\boldsymbol{R}$, as well as the prior distribution $p(\min[\text{eigen}(\boldsymbol{R})])$ of the minimum eigenvalue of $\boldsymbol{R}$ for the case with $K = 20$ corresponding to the value chosen in our empirical study in Subsection 4.2. By varying ν from 21 to 30, we observe a considerable effect of ν. Choosing $\nu = 25$, as we will do in Subsection 4.2, bounds the prior sufficiently away from regions where $\boldsymbol{R}$ is rank deficient and hence violates the identifiability conditions provided by Theorem 1.

It should be emphasized once more, that whether $\boldsymbol{S}$ is random as in the prior suggested by Huang and Wand (2013), or fixed, does not change the prior $p(\boldsymbol{R})$, leaving Bayesian inference invariant to this prior. However, it turns out that the prior of $\boldsymbol{S}$ influences the efficiency of the marginal data augmentation algorithm we use for inference, see Subsection 3.2.1, and mixing improves when $\boldsymbol{S}$ is random rather than fixed.

Finally, the marginal data augmentation algorithm will require sampling $\boldsymbol{\Lambda}$ from the conditional distribution $p(\boldsymbol{\Lambda} \mid \boldsymbol{R})$ for a given value of $\boldsymbol{R}$. Under the random prior for $\boldsymbol{S}$, we sample from the joint prior $p(\boldsymbol{\Lambda}, \boldsymbol{S} \mid \boldsymbol{R}) = p(\boldsymbol{\Lambda} \mid \boldsymbol{S}, \boldsymbol{R})\, p(\boldsymbol{S} \mid \boldsymbol{R}) = p(\boldsymbol{\Lambda} \mid \boldsymbol{S}, \boldsymbol{R})\, p(\boldsymbol{S})$, where $p(\boldsymbol{S} \mid \boldsymbol{R}) = p(\boldsymbol{S})$ is equal to the prior of $\boldsymbol{S}$, and the conditional distribution of $\boldsymbol{\Lambda} \mid \boldsymbol{S}, \boldsymbol{R}$ can be deduced from equation (19) using $p(\boldsymbol{\Lambda} \mid \boldsymbol{R}, \boldsymbol{S}) = p(\boldsymbol{\Lambda}, \boldsymbol{R} \mid \boldsymbol{S})/p(\boldsymbol{R} \mid \boldsymbol{S}) = p(\boldsymbol{\Lambda}, \boldsymbol{R} \mid \boldsymbol{S})/p(\boldsymbol{R})$.

Figure 1: Marginal prior distributions of the maximum correlation in absolute value ($p(\max_{j\neq k}|R_{jk}|)$, left panel) and of the smallest eigenvalue ($\min[\text{eigen}(\boldsymbol{R})]$, right panel) of the correlation matrix $\boldsymbol{R}$ in a model with $K = 20$, for different degrees of freedom for $\boldsymbol{R}$.



Notes. Kernel density estimation based on 10^5 draws from the prior distribution of $\boldsymbol{R}$.

It can be shown that each single variance $\Lambda_k \mid s_k, \boldsymbol{R}$ follows an inverse-Gamma distribution with s_k being drawn from the prior, i.e.:

$$s_k \sim \mathcal{G}\left(\frac{1}{2};\ \frac{1}{2\,\nu^\star A_k^2}\right), \qquad \Lambda_k \mid \boldsymbol{R}, s_k \sim \mathcal{G}^{-1}\left(\frac{\nu}{2};\ \frac{s_k r^{kk}}{2}\right). \tag{22}$$

If the scale matrix $\boldsymbol{S}$ is fixed, then $\Lambda_k \mid \boldsymbol{R}, s_k$ is sampled conditional on that value.

3 Bayesian Inference

Our inference approach is fully Bayesian and combines the likelihood function derived from model specification (1) under the assumptions on the latent factors θ_i and on the error terms ε_i specified in Subsection 2.2 with the prior distributions formulated in Subsection 2.3.

Our model contains a particular combination of ingredients (dedicated and correlated factors, dimension-varying structure, identification constraints) that requires a new procedure for Bayesian inference, based on Markov chain Monte Carlo (MCMC) methods.

For the fully specified model we consider in the present paper, the identification conditions formulated in Theorem 1 guarantee identifiability of $\boldsymbol{\Phi} = \{\boldsymbol{\Delta}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{R}\}$[17] in the classical

[17] $\boldsymbol{\Delta}$ is the matrix of binary indicators with the same dimensions as the factor loading matrix $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ is the matrix of regression coefficients capturing the effects of the covariates on the latent variables (see equation (1)), $\boldsymbol{\Sigma}$ are the idiosyncratic variances (see equation (3)), and $\boldsymbol{R}$ is the correlation matrix of the factors.

sense that any two solutions $\boldsymbol{\Phi}$ and $\boldsymbol{\Phi}'$ yielding the same likelihood for all possible realizations $\boldsymbol{Y}$, i.e., $p(\boldsymbol{Y} \mid \boldsymbol{\Phi}) = (\boldsymbol{Y} \mid \boldsymbol{\Phi}')$, are identical up to column and sign switching.

Within a Bayesian framework, the issue of identifiability is, in general, much less relevant. Any proper prior $p(\boldsymbol{\Phi})$ will turn a well-specified likelihood function $p(\boldsymbol{Y} \mid \boldsymbol{\Phi})$ into a proper posterior distribution $p(\boldsymbol{\Phi} \mid \boldsymbol{Y})$ by means of Bayes' theorem, $p(\boldsymbol{\Phi} \mid \boldsymbol{Y}) \propto p(\boldsymbol{Y} \mid \boldsymbol{\Phi})\, p(\boldsymbol{\Phi})$, even if positive prior probability is assigned to subspaces of the parameter space containing solutions that are not identified in the classical sense defined above. However, when it comes to practical Bayesian inference, such a posterior distribution does not necessarily lead to sensible estimates of the unknown parameters, if inference is based on averages of MCMC draws from the posterior distribution. To avoid the ambiguity inherent in a posterior distribution derived from the likelihood of an unidentified model, we pursue a more rigorous approach in the present paper and constrain the posterior $p(\boldsymbol{\Phi} \mid \boldsymbol{Y})$, by assigning positive prior probability $p(\boldsymbol{\Phi})$ only to parameters $\boldsymbol{\Phi}$ that are identified in the classical sense defined above.

Several computational challenges have to be overcome in implementing this approach. First, we develop a new search procedure to select the dimension and the structure of the latent part of the model, without jeopardizing the identification condition (Subsection 3.1). Second, allowing for correlated factors calls for a new sampling scheme of the correlation matrix in a dimension-varying model (Subsection 3.2).

3.1 MCMC Sampling Scheme to Produce Identified Models

Implementing the classical identifying conditions regarding the minimum number of measurements dedicated to each factor in equation (6) introduce nonstandard difficulties in a MCMC sampling scheme. To address this problem, we develop a new algorithm that produces classically identified models.

To extract meaningful factors and factor loadings from model (1), a value has to be assigned to the indicator matrix $\boldsymbol{\Delta}$. Different approaches have been proposed in the literature to estimate dimension-varying models. The most popular is the reversible jump Markov chain Monte Carlo (RJMCMC) algorithm of Green (1995), which can be designed to visit models of different dimensions during sampling. However, this sampler has some limitations. First, it requires that the analyst specifies alternative models to be compared in the algorithm. When there is no *a priori* knowledge about the structure of the factor loading matrix, nor about the number of factors, the number of potential models underlying the data is prohibitively large. Our Bayesian search procedure operates on the set of all possible matrices $\boldsymbol{\Delta}$, among the $(M \times K)$–dimensional indicator matrices belonging to the identified set $\mathcal{D}$, and allows us

to choose its value from the data. Second, RJMCMC requires running preliminary analyses for each of the alternative models to generate sensible proposal distributions (Lopes and West, 2004), which can be computationally very demanding and therefore impractical for application to large models.

To remedy these problems, alternative approaches relying on the Metropolis-Hastings algorithm (henceforth M-H, see Hastings, 1970; Chib and Greenberg, 1995) have been proposed. Borrowing from the literature on mixture modeling, the M-H sampler can, for instance, be tailored to implement dimension-changing moves that, at each MCMC iteration, attempt to merge some existing factors to shrink the dimension of the model, or, on the contrary, to split some existing factors to expand the model ("split & merge moves," see Richardson and Green, 1997). Alternatively, the sampler can attempt to introduce new factors sampled from their prior distribution, or to delete existing factors at each MCMC step ("birth & death moves," see Stephens, 2000). Again, the major difficulty with these approaches in large models is finding appropriate proposal distributions that will generate candidates for the split/merge or birth/death moves that are likely to be accepted as identified models.

The identifying requirements of our model (more specifically, the need to have at least three measurements dedicated to each factor), along with the specification of correlated factors, create nonstandard difficulties and prevent most MCMC algorithms from moving quickly enough through the parameter space to reach the stationary distribution of the parameters. This is a well-known issue in MCMC sampling. Recently, new approaches based on marginal data augmentation have been developed to handle these problems. These methods will be introduced in Subsection 3.2.1 for the sampling of the correlation matrix of the factors, but it is worth pointing out the analogy between our sampling scheme for the factor selection and marginal data augmentation methods. Both rely on intermediate steps in nonidentified models to boost the sampler, and both make sure that the algorithm always comes back to an identified model after these intermediate steps. But our approach differs in the sense that it does not introduce additional parameters into the model for this purpose, but rather relaxes restrictions on some existing parameters. More precisely, MCMC sweeps are carried out in the unrestricted version of the model (Subsection 3.1.1) to generate appropriate proposals for the M-H algorithm that will in the end only generate identified models (Subsection 3.1.2).

3.1.1 MCMC Sweeps in the Unrestricted Model

The MCMC sampler we implement to generate proposals draws model parameters and latent variables sequentially from their posterior distributions, conditioning at each step on the most recently drawn values of the other parameters and latent variables:

Algorithm 1 (**Unrestricted MCMC Sampler**). *The following steps are performed on the unrestricted model, i.e., where the constraint of at least three measurements dedicated to each factor is not enforced. The conditioning on the covariates $\boldsymbol{X}$ is implicitly assumed at each step:*

(A) *Sample the indicators $\boldsymbol{\Delta}$, the idiosyncratic variances $\boldsymbol{\Sigma}$ and the factor loadings $\boldsymbol{\alpha}$ simultaneously. Since $p(\boldsymbol{\alpha}, \boldsymbol{\Sigma}, \boldsymbol{\Delta} \mid \boldsymbol{Y}^\star, \boldsymbol{\theta}, \boldsymbol{\beta}, \tau) = p(\boldsymbol{\alpha} \mid \boldsymbol{Y}^\star, \boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\Delta}) p(\boldsymbol{\Sigma} \mid \boldsymbol{Y}^\star, \boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\Delta})$ $p(\boldsymbol{\Delta} \mid \boldsymbol{Y}^\star, \boldsymbol{\theta}, \boldsymbol{\beta}, \tau)$, this step can be broken down as follows:*

- ***(A-1)*** *Marginalize the distribution of $\boldsymbol{\Delta}$ with respect to $\boldsymbol{\Sigma}$ and $\boldsymbol{\alpha}$ and sample $\boldsymbol{\Delta}$ from $p(\boldsymbol{\Delta} \mid \boldsymbol{Y}^\star, \boldsymbol{\theta}, \boldsymbol{\beta}, \tau)$. Set the factor loadings corresponding to the zero indicators of $\boldsymbol{\Delta}$ to 0, and denote the remaining non-zero loadings as $\alpha^{\boldsymbol{\Delta}}$.*
- ***(A-2)*** *Marginalize the distribution of $\boldsymbol{\Sigma}$ with respect to $\alpha^{\boldsymbol{\Delta}}$ and sample $\boldsymbol{\Sigma}$ from $p(\boldsymbol{\Sigma} \mid \boldsymbol{Y}^\star, \boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\Delta})$.*
- ***(A-3)*** *Sample the non-zero factor loadings $\alpha^{\boldsymbol{\Delta}}$ from $p(\alpha^{\boldsymbol{\Delta}} \mid \boldsymbol{Y}^\star, \boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\Delta})$.*

(B) *Sample the regression coefficients $\boldsymbol{\beta}$ from $p(\boldsymbol{\beta} \mid \boldsymbol{Y}^\star, \boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\Sigma})$.*

(C) *For each binary measurement Y_{im}, sample the corresponding latent variable $Y_{im}^\star$ from $p(Y_{im}^\star \mid Y_{im}, \boldsymbol{\theta}, \beta_m, \boldsymbol{\alpha}_m)$, for $i = 1, \ldots, N$.*

(D) *Sample the factors $\boldsymbol{\theta}$ and their correlation matrix $\boldsymbol{R}$ jointly from $p(\boldsymbol{\theta}, \boldsymbol{R} \mid \boldsymbol{Y}^\star, \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\Sigma})$.*

(E) *Sample the indicator probabilities τ from $p(\tau \mid \boldsymbol{\Delta})$, or skip this step if τ is integrated out of the likelihood when the indicators are updated at step (A-1).*

Full details about the conditional distributions are provided in the subsequent sections and in Appendix A. Running this MCMC sampler on our factor model, where the indicators are sampled sequentially from their full conditional distributions, exhibits a good mixing of the Markov chain. There is, however, a major problem with this procedure, as it is not possible to force the algorithm to produce at least three measurements dedicated to each factor. As a consequence, this MCMC sampling scheme cannot be implemented to sample models that meet our identifiability requirements. We can nevertheless exploit these good

properties to generate relevant proposals, and embed these unrestricted MCMC sweeps into a M-H algorithm to construct a valid MCMC sampling scheme that produces identified models.

3.1.2 Metropolis-Hastings Moves to Produce Identified Models

The mechanics of our algorithm can be described as follows: at each MCMC iteration, a few unrestricted MCMC sweeps are performed to sample models where the number of measurements dedicated to each factor is not restricted. These intermediate steps can generate models that are nonidentified. The nonidentified samples, however, are not saved for posterior inference and only serve the purpose of visiting models of different dimensions to generate relevant proposals for the M-H moves. When navigating through (possibly) nonidentified models, not only the indicators $\boldsymbol{\Delta}$ are updated, but so are all of the parameters and latent variables of the system, in order to adjust all the components of the model. In so doing, the algorithm is more likely to reach an alternative state, where the factor loading matrix has a different structure (e.g., a different number of factors). New factors can, for instance, be introduced progressively into the model, one measurement at a time. The flexibility of the algorithm is the key to exploring models of different dimensions and finding the latent structure that is the most representative of the data.

The procedure can be summarized by the following algorithm:

Algorithm 2 (**M-H moves with intermediate steps in nonidentified models**). *Let* $\boldsymbol{\vartheta} = \{\boldsymbol{Y}^\star, \boldsymbol{\theta}, \boldsymbol{\Delta}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{R}, \tau\}$ *denote the set of model parameters and latent variables to be sampled. At each MCMC iteration, allow the Markov chain to temporarily visit nonidentified states of the model with unrestricted MCMC sweeps to generate a candidate that will be accepted (or rejected) by a M-H step. If the algorithm is currently in state* $\widehat{\boldsymbol{\vartheta}}_0$*, a candidate* $\check{\boldsymbol{\vartheta}}_0$ *is generated as follows by running* $2S$ *intermediate MCMC sweeps based on Algorithm 1:*

(M1) *Starting from* $\widehat{\boldsymbol{\vartheta}}_0$*, run* S *sweeps of the unrestricted MCMC sampler, by applying steps (A) to (E) iteratively, to produce a sequence* $\widehat{\boldsymbol{\vartheta}}_1, \widehat{\boldsymbol{\vartheta}}_2, \ldots, \widehat{\boldsymbol{\vartheta}}_{S-1}, \widehat{\boldsymbol{\vartheta}}_S =: \overline{\boldsymbol{\vartheta}}_S$*.*

(M2) *Starting from* $\check{\boldsymbol{\vartheta}}_S := \overline{\boldsymbol{\vartheta}}_S$*, run* S *sweeps of the unrestricted MCMC sampler* in reverse order *to produce a sequence* $\check{\boldsymbol{\vartheta}}_{S-1}, \ldots, \check{\boldsymbol{\vartheta}}_1, \check{\boldsymbol{\vartheta}}_0$*. Reverse moves are simply performed by drawing the parameters and latent variables in reverse order, i.e., from step (E) to step (A).*[18]

[18]Note that steps (A-1) to (A-3) are still performed in this order in the reverse move. Since they rely on the marginalization of some parameters, they cannot be performed in reverse order (van Dyk and Park, 2008). This is, however, not in contradiction with our approach, because only step (A) as a whole is relevant here, the sub-steps being only used to break it down into several pieces that are easier to perform separately. The complete MCMC sequence in reverse order therefore is: (E), (D), (C), (B), (A-1), (A-2), (A-3).

(M3) *Accept the candidate* $\check{\boldsymbol{\vartheta}}_0$ *as the new state if the resulting model is identified (i.e., if the corresponding* $\check{\boldsymbol{\Delta}}_0 \in \mathcal{D}$*), otherwise reject it and recycle the old state* $\widehat{\boldsymbol{\vartheta}}_0$ *as the new state of the Markov chain.*

The number $2S$ of intermediate steps is a tuning parameter that can be fixed *a priori*, or specified as stochastic (see Subsection 3.3 for more details). At this point, it remains to justify that the resulting Markov chain is valid, in the sense that it meets the minimum requirements ensuring that it converges to its stationary distribution. We now explain the intuition behind the theoretical foundations of our approach, and show that our algorithm satisfies the detailed balance condition.

Transition kernel and detailed balance condition. Let $p_u(\boldsymbol{\vartheta})$ denote the stationary distribution of $\boldsymbol{\vartheta}$ in the unrestricted model. For a transition kernel $T_u(\cdot, \cdot)$ associated with $p_u(\cdot)$, the detailed balance condition is verified if:

$$p_u(\widehat{\boldsymbol{\vartheta}})\, T_u(\widehat{\boldsymbol{\vartheta}}, \check{\boldsymbol{\vartheta}}) = p_u(\check{\boldsymbol{\vartheta}})\, T_u(\check{\boldsymbol{\vartheta}}, \widehat{\boldsymbol{\vartheta}}). \tag{23}$$

This condition is not necessary but is sufficient to show that $p_u(\cdot)$ is a stationary measure associated with the transition kernel T_u. It implies that the chain is reversible, i.e., that the probability of being in $\widehat{\boldsymbol{\vartheta}}$ and moving to $\check{\boldsymbol{\vartheta}}$ is the same as the probability of being in $\check{\boldsymbol{\vartheta}}$ and moving back to $\widehat{\boldsymbol{\vartheta}}$ (Casella and Robert, 2004, definition 6.45).

In the case where the transition is made of several sub-transitions applied sequentially, like in our unrestricted MCMC sampler, the transition kernel from a state $\widehat{\boldsymbol{\vartheta}}$ to a new state $\check{\boldsymbol{\vartheta}}$ through steps (A) to (E) is the product of the corresponding sub-transition kernels:

$$\begin{aligned} T_u(\widehat{\boldsymbol{\vartheta}}, \check{\boldsymbol{\vartheta}}) = p_u(\check{\boldsymbol{\alpha}}, \check{\boldsymbol{\Sigma}}, \check{\boldsymbol{\Delta}} \mid \widehat{\boldsymbol{Y}}^\star, \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\beta}}, \widehat{\tau})\, p(\check{\boldsymbol{\beta}} \mid \widehat{\boldsymbol{Y}}^\star, \widehat{\boldsymbol{\theta}}, \check{\boldsymbol{\alpha}}, \check{\boldsymbol{\Sigma}}) \\ \times p(\check{\boldsymbol{Y}}^\star \mid \boldsymbol{Y}, \widehat{\boldsymbol{\theta}}, \check{\boldsymbol{\beta}}, \check{\boldsymbol{\alpha}})\, p(\check{\boldsymbol{\theta}}, \check{\boldsymbol{R}} \mid \check{\boldsymbol{Y}}^\star, \check{\boldsymbol{\beta}}, \check{\boldsymbol{\alpha}}, \check{\boldsymbol{\Sigma}})\, p(\check{\tau} \mid \check{\boldsymbol{\Delta}}). \end{aligned}$$

Similarly, the transition kernel from $\check{\boldsymbol{\vartheta}}$ to $\widehat{\boldsymbol{\vartheta}}$ in reverse order, from step (E) to step (A), is:

$$\begin{aligned} T_u(\check{\boldsymbol{\vartheta}}, \widehat{\boldsymbol{\vartheta}}) = p(\widehat{\tau} \mid \check{\boldsymbol{\Delta}})\, p(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{R}} \mid \check{\boldsymbol{Y}}^\star, \check{\boldsymbol{\beta}}, \check{\boldsymbol{\alpha}}, \check{\boldsymbol{\Sigma}})\, p(\widehat{\boldsymbol{Y}}^\star \mid \boldsymbol{Y}, \widehat{\boldsymbol{\theta}}, \check{\boldsymbol{\beta}}, \check{\boldsymbol{\alpha}}) \\ \times p(\widehat{\boldsymbol{\beta}} \mid \widehat{\boldsymbol{Y}}^\star, \widehat{\boldsymbol{\theta}}, \check{\boldsymbol{\alpha}}, \check{\boldsymbol{\Sigma}})\, p_u(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\Sigma}}, \widehat{\boldsymbol{\Delta}} \mid \widehat{\boldsymbol{Y}}^\star, \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\beta}}, \widehat{\tau}). \end{aligned}$$

The detailed balance condition implies that both $T_u(\widehat{\boldsymbol{\vartheta}}, \check{\boldsymbol{\vartheta}})$ and the reverse move $T_u(\check{\boldsymbol{\vartheta}}, \widehat{\boldsymbol{\vartheta}})$ have $p_u(\cdot)$ as stationary distribution. Nevertheless, $p_u(\boldsymbol{\vartheta})$ is not our targeted distribution, as it can generate nonidentified models. Rather, we are looking for a stationary distribution $p(\boldsymbol{\vartheta})$ on the set of identified models that also verifies the detailed balance condition in

equation (23), i.e.,

$$p(\boldsymbol{\vartheta}) \propto p_u(\boldsymbol{\vartheta})\, \delta_{\mathcal{D}}(\boldsymbol{\Delta}), \tag{24}$$

where $\boldsymbol{\Delta} \in \mathcal{D}$, and $\delta_{\mathcal{D}}(\boldsymbol{\Delta})$ is the Dirac measure that is equal to 1 if $\boldsymbol{\Delta} \in \mathcal{D}$, to 0 otherwise.

A parallel can be drawn between our method relying on intermediate steps in unrestricted models and Neal (1996)'s tempered transitions, which are designed as a very general approach to sample from multimodal distributions.[19] Nevertheless, it should be emphasized that our approach departs from Neal (1996), as we relax the identifying restrictions during the intermediate steps, while the tempered transitions always operate on identified models. This is a major difference between the two approaches. The proof of the detailed balance condition, however, looks very similar. We present it in Appendix A.1 for the sake of completeness.

The symmetry of the intermediate moves aids in simplifying computations, as it bypasses the need to calculate the normalizing constant in equation (16). This results in a very simple form for the acceptance rate: proposed $\check{\boldsymbol{\vartheta}}_0$ are automatically accepted as a new state of the model if their corresponding indicator matrix $\check{\boldsymbol{\Delta}}_0$ belongs to the identified set, otherwise they are rejected.

The MCMC sweeps performed to sample the parameters and the latent variables of the model are straightforward to implement, except for the correlation matrix of the latent factors, which requires some elaboration. We now discuss this specific stage, and explain the technical improvements of our sampling scheme over previous algorithms.

3.2 Sampling the Latent Factors and their Correlation Matrix in a Dimension-Varying Model

Ours is the first paper in the Bayesian factor analysis literature to consider correlated factors in a dimension-varying model where identification of the model is secured explicitly. This feature of the model is challenging for the sampling procedure in two respects. First, drawing a correlation matrix is not trivial, because of the combination of fixed diagonal elements and positive-definiteness. Since no natural conjugate distribution exists for this matrix, the usual Gibbs sampler cannot be implemented. Subsection 3.2.1 discusses this issue and presents the approach we adopt that relies on marginal data augmentation. Second, the dimension of the latent part of our model is not fixed and varies during sampling. This implies that correlation matrices of different sizes, dependent on the number of latent factors, have to

[19] The tempered transitions are performed through the use of a sequence of intermediate distributions that are "heated" by different temperature parameters to flatten the likelihood function, thus allowing bigger moves.

be sampled through MCMC iterations. Subsection 3.2.2 introduces the block sampling we develop to cope with this problem.

3.2.1 Sampling the Correlation Matrix through Marginal Data Augmentation

We borrow from the literature on marginal data augmentation to sample the correlation matrix of the factors and to boost the MCMC sampling of the factor loadings and of the factors at the same time. To the best of our knowledge, this simple idea has not been applied to factor models with correlated factors in the literature.

Marginal data augmentation (henceforth MDA, see Meng and van Dyk, 1999; van Dyk and Meng, 2001; Imai and van Dyk, 2005), also referred to as *parameter-expanded data augmentation* (Liu and Wu, 1999), has recently been proposed as a very general and simple way to improve the convergence and the mixing of Markov chains in MCMC sampling. We apply this approach to achieve this primary goal of boosting convergence and mixing, but also, and maybe more importantly, to develop a new sampling scheme for the correlation matrix that turns out to be easier to implement than existing methods based on the M-H algorithm (Zhang et al., 2006; Liu and Daniels, 2006; Liu, 2008).

MDA consists of expanding the parameter space, at each MCMC iteration, by introducing a set of parameters that do not belong to the original model, and that usually cannot be identified from the data. Once the model has been transformed appropriately with these so-called "working parameters," a Gibbs sweep is performed in the expanded model (which is usually easier to perform than in the original model), and the model is finally transformed back to its original form. It is important to note that this expansion of the model is temporary and is only used as a computational device. The draws produced in the expanded model are not saved for posterior inference. Only the values of the parameters resulting from the final transformation are saved.

In our factor model, the variances of the factors are restricted to 1 for purposes of identification. This restriction can easily be relaxed to expand the model, using these variances as working parameters. Assume for now that the dimension of the model is fixed at K factors, and that we are therefore sampling a correlation matrix $\boldsymbol{R}$ of dimension $(K \times K)$ in the original model, and a covariance matrix $\boldsymbol{\Omega} = \boldsymbol{\Lambda}^{\frac{1}{2}} \boldsymbol{R} \boldsymbol{\Lambda}^{\frac{1}{2}}$, where $\boldsymbol{\Lambda} = \text{diag}(\Lambda_1, \ldots, \Lambda_K)$, of same dimensions in the expanded model. At a given MCMC iteration (t), MDA proceeds as follows when it comes to the update of $\boldsymbol{R}$:

- **Model Augmentation.** Expand the model with the variances of the factors $\boldsymbol{\Lambda}$ used as working parameters. Since no information is available about these parameters conditional on $\boldsymbol{R}^{(t-1)}$, they are sampled from the prior distribution $p(\boldsymbol{\Lambda} \mid \boldsymbol{R}^{(t-1)})$ according to

equation (22), where the current value $r^{kk(t-1)}$ is used to sample each Λ_k, $k = 1, \ldots, K$ conditional on a scale matrix $\boldsymbol{S}_{\text{prior}}^{(t)}$ drawn from the prior $p(\boldsymbol{S})$. Call this draw $\boldsymbol{\Lambda}_{\text{prior}}^{(t)}$, and transform the model as follows, for $i = 1, \ldots, N$:[20]

$$\widetilde{\boldsymbol{\alpha}}^{(t)} = \boldsymbol{\alpha}^{(t)} \left(\boldsymbol{\Lambda}_{\text{prior}}^{(t)}\right)^{-\frac{1}{2}}, \qquad \widetilde{\boldsymbol{\theta}}^{(t)} = \boldsymbol{\theta}^{(t)} \left(\boldsymbol{\Lambda}_{\text{prior}}^{(t)}\right)^{\frac{1}{2}},$$

so that in the expanded model $\widetilde{\theta}_i^{(t)} \sim \mathcal{N}\left(0;\, \widetilde{\boldsymbol{\Omega}}^{(t)}\right)$ with

$$\widetilde{\boldsymbol{\Omega}}^{(t)} = \left(\boldsymbol{\Lambda}_{\text{prior}}^{(t)}\right)^{\frac{1}{2}} \boldsymbol{R}^{(t-1)} \left(\boldsymbol{\Lambda}_{\text{prior}}^{(t)}\right)^{\frac{1}{2}}.$$

- **Update the covariance matrix in the expanded model** using a Gibbs step:

$$\boldsymbol{\Omega}^{(t)} \mid \boldsymbol{S} \sim \mathcal{W}_K^{-1}\left(\nu + N;\, \boldsymbol{S} + \widetilde{\boldsymbol{\theta}}^{(t)\prime}\widetilde{\boldsymbol{\theta}}^{(t)}\right),$$

to obtain the updated working parameters $\boldsymbol{\Lambda}_{\text{post}}^{(t)}$ from the diagonal of $\boldsymbol{\Omega}^{(t)}$.

Concerning the scale matrix $\boldsymbol{S}$ applied in this step, it could be set equal to the scale matrix sampled from the prior, i.e., $\boldsymbol{S} = \boldsymbol{S}_{\text{prior}}^{(t)}$. Alternatively, $\boldsymbol{S}$ could be updated prior to sampling $\boldsymbol{\Omega}^{(t)}$ by sampling $\boldsymbol{S}^{(t)}$ from $p(\boldsymbol{S} \mid \boldsymbol{\Omega})$ conditional on $\boldsymbol{\Omega} = \widetilde{\boldsymbol{\Omega}}^{(t)}$. The corresponding posterior $p(\boldsymbol{S} \mid \boldsymbol{\Omega})$ is easily derived,

$$\begin{aligned} p(\boldsymbol{S} \mid \boldsymbol{\Omega}) &\propto p(\boldsymbol{\Omega} \mid \boldsymbol{S}) p(\boldsymbol{S}), \\ &\propto |\boldsymbol{S}|^{\frac{\nu}{2}} \exp\left\{-\frac{1}{2}\operatorname{tr}(\boldsymbol{S}\boldsymbol{\Omega}^{-1})\right\} \prod_{k=1}^{K} s_k^{-1/2} \exp\left\{-\frac{s_k}{2A_k^2(\nu - K + 1)}\right\}, \\ &\propto \prod_{k=1}^{K} (s_k)^{\frac{\nu+1}{2}-1} \exp\left\{-\frac{s_k}{2}\left((\boldsymbol{\Omega}^{-1})_{kk} + \frac{1}{A_k^2(\nu - K + 1)}\right)\right\}, \end{aligned}$$

and yields

$$s_k \mid \boldsymbol{\Omega} \sim \mathcal{G}\left(\frac{\nu+1}{2};\, \frac{(\boldsymbol{\Omega}^{-1})_{kk} + [A_k^2(\nu - K + 1)]^{-1}}{2}\right).$$

- **Transform back to the identified model:**

$$\begin{aligned} \boldsymbol{\alpha}^{(t)} &\longleftarrow \widetilde{\boldsymbol{\alpha}}^{(t)} \left(\boldsymbol{\Lambda}_{\text{post}}^{(t)}\right)^{\frac{1}{2}}, \qquad & \boldsymbol{R}^{(t)} &= \left(\boldsymbol{\Lambda}_{\text{post}}^{(t)}\right)^{-\frac{1}{2}} \boldsymbol{\Omega}^{(t)} \left(\boldsymbol{\Lambda}_{\text{post}}^{(t)}\right)^{-\frac{1}{2}}, \\ \boldsymbol{\theta}^{(t)} &\longleftarrow \widetilde{\boldsymbol{\theta}}^{(t)} \left(\boldsymbol{\Lambda}_{\text{post}}^{(t)}\right)^{-\frac{1}{2}}, & \end{aligned}$$

[20] Here it is assumed that $\boldsymbol{\alpha}$ and $\boldsymbol{\theta}$ have already been updated in the current MCMC iteration, hence their superscript (t); $\boldsymbol{\alpha}$ is the factor loading matrix, see equation (1).

where the left arrows ($\longleftarrow$) indicate that the current values of the factor loadings and of the latent factors at iteration (t) are replaced by the corresponding transformed values. Note that this backward transformation is deterministic, given the updated variances $\boldsymbol{\Lambda}_{\text{post}}^{(t)}$.

These transformations are the mechanism of the marginal data augmentation that allows the sampling of the correlation matrix, improving the mixing of the Markov chain at the same time.

3.2.2 Block Sampling of the Correlation Matrix Jointly with the Factors

We specify a maximum number K of factors *a priori*, but not all of them will ultimately be loaded by measurements.[21] We make a distinction between the factors that have an impact on the measurements and belong to the identifiable set (those loaded by at least three measurements, called "active" factors) and those that do not (the "inactive" factors, which are not loaded by any measurements). The former correspond to the non-zero columns of the factor loading matrix $\boldsymbol{\alpha}$, and the latter to the zero columns. The inactive factors can be regarded as potential new factors, as it can happen, at any time during sampling, that some measurements start loading on them. Conversely, existing (active) factors can be shut down and become inactive if their dedicated measurements no longer load on them at a given MCMC iteration.

Assume that at a particular stage there are K_1 active factors and K_2 inactive factors, with $K_1 + K_2 = K$. The latent factors are reordered such that the K_1 active factors ($\boldsymbol{\theta}_1$) appear first and the K_2 inactive factors ($\boldsymbol{\theta}_2$) appear in the last positions of $\boldsymbol{\theta}$. The rows and/or columns of the different parameters and latent variables are thus reordered and partitioned as follows:

$$\boldsymbol{\theta} = \begin{pmatrix} \boldsymbol{\theta}_1 & \boldsymbol{\theta}_2 \end{pmatrix}, \quad \boldsymbol{\alpha} = \begin{pmatrix} \boldsymbol{\alpha}_1 & \boldsymbol{\alpha}_2 \end{pmatrix}, \quad \boldsymbol{R} = \begin{pmatrix} \boldsymbol{R}_{11} & \boldsymbol{R}_{12} \\ \boldsymbol{R}_{21} & \boldsymbol{R}_{22} \end{pmatrix}, \quad \boldsymbol{\Omega} = \begin{pmatrix} \boldsymbol{\Omega}_{11} & \boldsymbol{\Omega}_{12} \\ \boldsymbol{\Omega}_{21} & \boldsymbol{\Omega}_{22} \end{pmatrix}, \tag{25}$$

where $\boldsymbol{R}$ is the correlation matrix of the factors, and $\boldsymbol{\Omega}$ is the corresponding covariance matrix (see equation (3)). A naive approach would be to sample the latent factors (active and inactive) and their correlation matrix sequentially through Gibbs sampling. However, mixing can be very poor in latent variable models. In our case, the draws of the correlations of the inactive factors would be highly autocorrelated across MCMC iterations if we sampled in this fashion. This would, in turn, affect the search procedure, as the sampled inactive

[21] If the sampler actually reaches the maximum number of factors K, the model should be reestimated with a larger value of $K \leq K^{\max}$, as more factors may be underlying the data.

factors—the potential new factors—would be very similar across MCMC iterations, making it difficult for the algorithm to pick new factors to better fit the data.

To remedy the slow mixing problem, the inactive factors and the covariance matrix $\boldsymbol{\Omega}$ are sampled simultaneously in the augmented model of the marginal data augmentation procedure. This blocking strategy has been shown to substantially improve mixing and convergence (Liu et al., 1994). The sampling procedure is carried out in two steps. First, since the likelihood does not depend on the inactive factors (since $\boldsymbol{\alpha}_2 = \mathbf{0}$), these factors $\boldsymbol{\theta}_2$ can be integrated out and the active factors can be updated marginally (van Dyk and Park, 2008). The marginal conditional prior distribution of θ_{1i} is $\mathcal{N}(0;\, \boldsymbol{\Omega}_{11})$, and the updated conditional posterior is derived as follows, for all $i = 1, \ldots, N$:[22]

$$\theta_{1i} \mid \boldsymbol{\Omega}_{11}, \ldots \sim \mathcal{N}(\boldsymbol{A}_{\theta_1} a_{\theta_{1i}};\, \boldsymbol{A}_{\theta_1})\,,$$

with:

$$(\boldsymbol{A}_{\theta_1})^{-1} = \boldsymbol{\alpha}_1{}'(\boldsymbol{\Sigma})^{-1}\boldsymbol{\alpha}_1 + (\boldsymbol{\Omega}_{11})^{-1}, \qquad a_{\theta_{1i}} = \boldsymbol{\alpha}_1{}'(\boldsymbol{\Sigma})^{-1}(Y_i^\star - \boldsymbol{\beta} X_i).$$

Once $\boldsymbol{\theta}_1$ has been updated, the inactive factors and the whole covariance matrix can be sampled simultaneously. Their joint distribution, in the expanded model, is proportional to:

$$p(\boldsymbol{\theta}_2, \boldsymbol{\Omega} \mid \boldsymbol{\theta}_1, \boldsymbol{Y}, \boldsymbol{\alpha}, \boldsymbol{\Sigma}) \propto p(\boldsymbol{\theta}_2 \mid \boldsymbol{\Omega}, \boldsymbol{\theta}_1)\, p(\boldsymbol{\Omega}_{12}, \boldsymbol{\Omega}_{22} \mid \boldsymbol{\Omega}_{11})\, p(\boldsymbol{\Omega}_{11} \mid \boldsymbol{\theta}_1),$$

revealing that the covariance matrix $\boldsymbol{\Omega}$ can be sampled by blocks. For this purpose, we develop a sampling procedure that relies on well-known properties of the inverse-Wishart distribution.[23] More precisely, we exploit the fact that the matrix $\boldsymbol{\Omega}_{11}$ is independent of the block matrices ($\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12}$ and $\boldsymbol{\Omega}_{22\cdot1}$), where $\boldsymbol{\Omega}_{22\cdot1} = \boldsymbol{\Omega}_{22} - \boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12}$ is the Schur complement of $\boldsymbol{\Omega}_{11}$ in $\boldsymbol{\Omega}$, both *a priori* as well as *a posteriori*. Hence, we split the scale matrix $\boldsymbol{S}$ appearing in the inverse Wishart prior and, respectively, posterior distribution of $\boldsymbol{\Omega}$ in a similar way as in equation (25):

$$\boldsymbol{S} = \begin{pmatrix} \boldsymbol{S}_{11} & \boldsymbol{S}_{12} \\ \boldsymbol{S}_{21} & \boldsymbol{S}_{22} \end{pmatrix}.$$

[22] To make the notation lighter, in this section we drop the tildes characterizing the transformed parameters of the MDA, although all these steps are carried out in the augmented model described in Subsection 3.2.1.

[23] See Theorem A1 in Web Appendix, available at http://heckman.uchicago.edu/BayesFA.

Using the prior $\boldsymbol{\Omega}_{11} \sim \mathcal{W}_{K_1}^{-1}(\nu - K_2;\, \boldsymbol{S}_{11})$, in a first step we sample the block matrix $\boldsymbol{\Omega}_{11}$ conditional on $\boldsymbol{\theta}_1$ from the posterior

$$\boldsymbol{\Omega}_{11} \mid \boldsymbol{\theta}_1 \sim \mathcal{W}_{K_1}^{-1}(\nu - K_2 + N;\, \boldsymbol{S}_{11} + \boldsymbol{\theta}_1'\boldsymbol{\theta}_1)\,.$$

Given the independence of the blocks stated above, in a second step, we sample the Schur complement $\boldsymbol{\Omega}_{22\cdot1}$ and the product $\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12}$ jointly:

$$\begin{aligned}
&\boldsymbol{\Omega}_{22\cdot1} \sim \mathcal{W}_{K_2}^{-1}(\nu;\, \boldsymbol{S}_{22\cdot1})\,,\\
&\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12} \mid \boldsymbol{\Omega}_{22\cdot1} \sim \mathcal{N}_{K_1\times K_2}\big(\boldsymbol{S}_{11}^{-1}\boldsymbol{S}_{12};\, \boldsymbol{S}_{11}^{-1} \otimes \boldsymbol{\Omega}_{22\cdot1}\big)\,.
\end{aligned}$$

Once these different blocks of the covariance matrix have been sampled, the inactive factors are sampled in a final step from the conditional distribution $p(\boldsymbol{\theta}_2 \mid \boldsymbol{\Omega}, \boldsymbol{\theta}_1)$ independently for all $i = 1, \ldots, N$:

$$\theta_{2i} \mid \boldsymbol{\Omega}, \theta_{1i} \sim \mathcal{N}\big((\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12})'\theta_{1i};\, \boldsymbol{\Omega}_{22\cdot1}\big)\,.$$

This block strategy of sampling the latent factors and their correlation matrix simultaneously dramatically improves the mixing of the algorithm and, in turn, facilitates factor selection.

3.3 Prior Specification and MCMC Tuning: Some Guidelines

The prior parameters should be carefully specified for the factor selection to work appropriately. Those discussed below, especially, play a crucial role and require particular attention.

The prior variance of the non-zero factor loadings defined in (18) is proportional to the idiosyncratic variance of each measurement, with a scale parameter A_m^0 that determines how diffuse the distribution is. Subsequently, we make use of a fixed scale prior, where $A_m^0 = A_0$. Although it is usually not recommended to specify vague priors in latent variable models (e.g., $A_0^{-1} = 0$), as the near impropriety of the resulting posterior distribution can lead to a slow mixing of the sampler (Natarajan and McCulloch, 1998), being too informative should also be avoided. Too small a scale parameter would shrink the distribution of the loadings toward 0, especially in cases where measurement error is small. This could in turn induce an overfitting of the number of latent factors, where many factor loadings would have a low magnitude.

The degrees of freedom ν in the prior of the covariance matrix of the factors in the expanded model defined in (19) determines the marginal prior distribution of the factor correlations. Taking $\nu = K + 1$ such that the single correlations are uniformly distributed

on $[-1; 1]$ (see Barnard et al., 2000) can be problematic in high-dimensional models. It may indeed result in an overestimation of the number of latent factors, where many factors would appear to be extremely highly correlated and therefore redundant to explain the data. To cope with this *factor splitting* problem, it might be helpful to increase ν to prevent duplicate factors from emerging. As outlined previously in Subsection 2.3.5 and at the beginning of this section, increasing ν is also important with respect to ensuring prior identification in cases where the likelihood function yields considerable support for unidentifiable regions of the parameter space.

The prior on the indicators' probabilities τ (see equation (13)) needs to be tailored appropriately for the factor selection process. Due to the identifying constraints on the indicator matrix $\boldsymbol{\Delta}$, the implied prior distribution on the number of factors appears to be very tedious to derive analytically. It can however easily be simulated. Table 2 shows the prior probabilities of the numbers of factors for some models studied in the Monte Carlo experiment.

When τ_0 is specified individually for each measurement (see Subsection 2.3.1), the impact of its prior specification vanishes if the Beta distribution is specified as symmetric (i.e., with equal shape parameters). This might appear counterintuitive at first sight, as one could expect a crucial role of the prior distributions of τ_0 in the Bayesian updating process when only one observation of Δ_m is available at each MCMC iteration. However, with a single observation at hand, only the mean of the prior distribution counts, and this one is not affected by a change of scale of the prior parameters.[24] This explains why there is no difference between using a uniform prior for τ_0 (i.e., $\mathcal{B}eta(1; 1)$) and a very informative prior such as a U-shaped distribution reflecting the belief that τ_0 is either close to 0 or to 1 (e.g., $\mathcal{B}eta(0.1; 0.1)$).

The number $2S$ of intermediate steps determines how long the algorithm navigates through expanded models to generate proposals for the M-H moves, and turns out to play an important role in the convergence of the algorithm. It can be specified as fixed or stochastic (e.g., sampled from a Poisson distribution at each MCMC iteration) to introduce more flexibility in the M-H algorithm. In some situations, for instance when the sampler is stuck in one region of the parameter space and does not move, additional intermediate steps can be performed to allow the algorithm to reach another state.

Convergence of the M-H algorithm can be slow in large models, due to the huge dimension of the parameter space. The choice of the initial value for the indicator matrix $\boldsymbol{\Delta}$ therefore

[24] The prior mean of τ_0 is $\kappa_0/(\kappa_0+\xi_0)$, and this ratio is not affected by a change of scale of the parameters, as long as these parameters remain proportional. This can also be seen from the ratio of the marginal likelihoods of $\boldsymbol{\Delta}$ in equation (A9), which remains the same after such a change of scale.

plays an important role. Instead of choosing this matrix at random, we suggest to run a preliminary MCMC analysis based on the unrestricted sampler (Algorithm 1) to generate an appropriate starting value. This sampler can be implemented to explore the parameter space more quickly, but it will generate a factor loading matrix that is only *partially identified*, in the sense that it contains columns with at least three non-zero factor loadings, but possibly also columns with less than three non-zero values. Such a partially identified matrix can however be used to generate a starting value for $\boldsymbol{\Delta}$ that corresponds to an identified model, by keeping only the columns with at least three non-zero values. The measurements dedicated to unidentified factors (with less than two dedicated measurements) can then be allocated either at random or according to our allocation rule to the identified factors. This approach based on the partial identification of the factor loading matrix can be theoretically justified (see, for instance, Sato, 1992, Theorem 3.9), and it can considerably reduce the need for a long burn-in period in practice.

3.4 Posterior Inference

The use of indicators makes it very easy to summarize the structure of the factor loading matrix. For example, the number D_k of measurements that are dedicated to a given factor k, for $k = 1, \ldots, K$, the number of discarded measurements D_0, the number of active factors K_1, or the number of included measurements $\widetilde{M}$ (those actually loading on a latent factor), can be computed as:

$$D_k = \sum_{m=1}^{M} \mathbf{1}[\Delta_m = e_k], \quad K_1 = \sum_{k=1}^{K} \mathbf{1}[D_k \neq 0], \quad \widetilde{M} = \sum_{k=1}^{K} D_k, \quad D_0 = M - \widetilde{M}.$$

These quantities can all be estimated using the corresponding posterior modes or posterior means over the MCMC draws, and are not affected by the column switching problem, nor by the sign switching problem. These two problems should, however, be dealt with (i.e., identification of the model should be restored *a posteriori*) to be able to interpret the latent structure of the factor loading matrix.

Since there is no natural ordering of the columns of the factor loading matrix, different approaches can be adopted to solve the column switching problem. We suggest a reordering based on the top elements of the columns, i.e., the first row l_k in each active column k containing a non-zero factor loading, starting from the top of the matrix. Because of the dedicated structure of the factor loading matrix, each of these top elements corresponds to a different measurement. At each MCMC iteration, the non-zero columns of $\boldsymbol{\alpha}$ are reordered such that the top elements appear in increasing order, i.e., $l_1 < l_2 < \ldots < l_K$. Finally,

the rows and columns of the correlation matrix $\boldsymbol{R}$ of the factors should also be switched accordingly.

Regarding the sign switching issue, a simple sign switch can be carried out on the MCMC draws to reestablish the consistency of the signs across iterations. To do so, one factor loading is used as a benchmark in each column (e.g., the factor loading with the highest posterior probability of being different from zero in each column[25]). The analyst determines which sign each benchmark loading should have, and the MCMC draws are then post-processed. Whenever the benchmark has the wrong sign in a certain column, sign switching has occurred at the corresponding MCMC iteration and is reversed by switching the signs of all the loadings that are in the same column (including the benchmark), of the latent factors corresponding to this column, as well as of the corresponding elements in the correlation matrix $\boldsymbol{R}$ of the factors.[26]

The decision on defining the signs of the loadings used as benchmarks should be guided by the meaning of the latent traits measured by the factors. If a factor captures a positive trait, like self-esteem, and the corresponding measurements are increasing in this trait, then it is straightforward to assume that the sign of the benchmark is positive, because a negative loading would capture the reverse of the trait of interest. The analyst should therefore always have the underlying literature in mind when carrying out this step, so as not to produce results that are counterintuitive and hard to interpret.

4 Applications to Simulated and Real Data

4.1 Monte Carlo Study

Data Generation. To investigate the performance of our algorithm, we run a Monte Carlo experiment using synthetic data simulated from a simplified version of equation (1). Since the focus of the experiment is on the factor selection process, no covariates are specified and the measurements are all assumed to be continuous (i.e., $Y_m = Y_m^\star$), so as to keep the specification as simple as possible.

We generate models of different dimensions and denote them by $\mathcal{M}(M, K_0, D, D_0)$, where M is the total number of measurements, K_0 the true number of factors, D the number of measurements dedicated to each factor, and D_0 the number of extra measurements that are uncorrelated with the other measurements.

[25] If several factor loadings have the same highest posterior probability (e.g., 1.0), we simply take the first of them from the top of the matrix.

[26] Frühwirth-Schnatter and Lopes (2012) use a similar approach to address the sign and column switching problems.

Each model is made of $M = K_0 D + D_0$ measurements that are dedicated to the latent factors through the following indicator matrix:

$$\underset{(M \times K_0)}{\boldsymbol{\Delta}} = \begin{bmatrix} \boldsymbol{I}_{K_0} \otimes \iota_D \\ \mathbf{0}_{(D_0 \times K_0)} \end{bmatrix},$$

where $\iota_D = (1, \ldots, 1)'$ is the vector of ones of length D. The uncorrelated measurements (if any) are placed at the bottom of the vector of measurements, hence the last D_0 zero rows of $\boldsymbol{\Delta}$. For the correlated measurements, each single non-zero factor loading $\alpha_m^{\boldsymbol{\Delta}}$ and each idiosyncratic variance σ_m^2 are simulated independently from the following distributions:

$$\begin{aligned} \alpha_m^{\boldsymbol{\Delta}} &= (-1)^{\phi_m} \sqrt{a_m}, & \sigma_m^2 &\sim \mathcal{U}(0.20;\, 0.80)\,, \\ \phi_m &\sim \mathcal{B}er\,(0.5)\,, & a_m &\sim \mathcal{U}(0.04;\, 0.64)\,, \end{aligned}$$

for $m = 1, \ldots, K_0 D$, where non-zero factor loadings $\alpha_m^{\boldsymbol{\Delta}}$ are assigned a sign at random with probability 0.5. The remaining D_0 uncorrelated measurements are simulated independently from a standard normal distribution, i.e., $\sigma_m^2 = 1$, for $m = K_0 D + 1, \ldots, M$, and the corresponding last rows of $\boldsymbol{\alpha}$ contain only zero elements. The correlation matrix $\boldsymbol{R}$ of the factors is sampled as

$$\boldsymbol{\Omega} \sim \mathcal{W}_{K_0}^{-1}(K_0 + 5;\, \boldsymbol{I}_{K_0})\,, \qquad \boldsymbol{\Lambda} = \operatorname{diag}(\boldsymbol{\Omega}), \qquad \boldsymbol{R} = \boldsymbol{\Lambda}^{-\frac{1}{2}}\, \boldsymbol{\Omega}\, \boldsymbol{\Lambda}^{-\frac{1}{2}},$$

where $\boldsymbol{\Omega}$ is the factor covariance matrix (see equation (3)), and the distribution of $\boldsymbol{R}$ is truncated to the subspace where all off-diagonal elements are smaller than 0.85 to avoid extreme cases.[27]

Model parameters are sampled independently across Monte Carlo replications. Drawing the factor loadings and the idiosyncratic variances from these uniform distributions results in measurements with a proportion of noise $\rho_m = \sigma_m^2/(\sigma_m^2 + a_m)$ that ranges from 24% to 95% for the correlated measurements. The signal-to-noise ratio is comparable to what we observe in our real data application. It is worth emphasizing that factor extraction is very challenging in this context of noisy data.[28]

We simulate the following eight models, where the number of measurements ranges from 15 to 125, and the number of factors from 3 to 12:

$$\mathcal{M}(15, 3, 5, 0), \qquad \mathcal{M}(36, 6, 6, 0), \qquad \mathcal{M}(72, 9, 8, 0), \qquad \mathcal{M}(120, 12, 10, 0),$$

[27] Thus, any simulated $\boldsymbol{R}$ with at least one correlation large than 0.85 is discarded and a new $\boldsymbol{R}$ is simulated. The operation is repeated until a correlation matrix satisfying this restriction is sampled.

[28] See Web Appendix for additional Monte Carlo experiments with less noisy data.

$$\mathcal{M}(17,3,5,2), \qquad \mathcal{M}(39,6,6,3), \qquad \mathcal{M}(76,9,8,4), \qquad \mathcal{M}(125,12,10,5).$$

Each of these model configurations is used to generate data sets with $N = 500$ and $1,000$ observations. For each of these data sets, 100 Monte Carlo replications are used.[29]

Prior Specification and MCMC Tuning. Table 1 displays the values of the prior parameters specified for this Monte Carlo study. For the correlation matrix of the factors, we implement the Huang-Wand prior by specifying a stochastic scale matrix $\boldsymbol{S}$ for the inverse-Wishart of $\boldsymbol{\Omega}$ that is updated at each MCMC iteration in the expanded model (see Subsection 3.2). The tuning parameter $\nu^\star = \nu - K + 1$ is chosen to induce a uniform prior distribution on $[-1;1]$ on the individual correlations of the factors. The prior on the indicator matrix is specified to allow uncorrelated measurements to be easily discarded from the model. Following Subsection 2.3.1, the probability of a zero row in the factor loading matrix is specified as measurement-specific. Conditional on the inclusion of the measurements into the model, the Dirichlet distribution on $\tau^\star$ is then specified differently for each model size, so as to generate plausible prior probabilities for the number of factors. Table 2 shows these prior probabilities for the first four models under investigation. These probabilities were simulated using a simple accept-reject sampling scheme and the low acceptance rates in the last column reflect the difficulty in sampling models that meet the identifying restrictions when drawing only from unrestricted models.

For each Monte Carlo replication, the MCMC sampler is run for a total of $40,000$ iterations, where only the last $20,000$ iterations are saved for posterior inference. The factor search is carried out with a number of $2S$ intermediate steps, where S is drawn randomly at each MCMC iteration as $S = 1 + \phi$, with $\phi \sim \mathcal{P}oisson\,(4)$.[30] The starting values of the parameters are selected at random, except for the indicator matrix $\boldsymbol{\Delta}$, which is specified after a pre-MCMC analysis. This preliminary analysis is performed by running the unrestricted

Table 1: Baseline Prior Specification for the Monte Carlo Study

Parameters	Values
Indicator matrix	$\kappa_0 = \xi_0 = 0.1$ and $\kappa = 1.0$ / 0.8 / 0.5 for $K_0 = 3$, 6 / 9 / 12
Idiosyncratic variances	$c_0 = 2.5$ and C_m^0 specified to avoid a Heywood problem
Factor loadings	$a_m^0 = 0$ and $A_m^0 = 3.0$
Factor correlation matrix	$\nu^\star = 2$ and $A_k^2 = 1/2$ (Huang-Wand prior)

[29] Therefore, this Monte Carlo experiment relies on 8 (model sizes) $\times$ 2 (sample sizes) $\times$ 100 (Monte Carlo replications) $= 1,600$ independent data sets.

[30] Which results in an average number of 10 steps in expanded models.

Table 2: Prior distribution of the number of factors induced by $\tau^\star \sim \mathcal{D}ir(\kappa, \ldots, \kappa)$ in models with M measurements and K potential factors, conditional on the inclusion of all measurements. True number of factors in Monte Carlo studies in bold.

M	K	κ	#Factors / Probability							Acc.
15	5	1.0		1	2	**3**	4	5		
				0.010	0.206	**0.576**	0.205	0.002		0.125
36	9	1.0	3	4	5	**6**	7	8	9	
			0.005	0.051	0.219	**0.390**	0.269	0.063	0.003	0.041
72	12	0.8	6	7	8	**9**	10	11	12	
			0.026	0.104	0.243	**0.316**	0.222	0.075	0.009	0.041
120	18	0.5	9	10	11	**12**	13	14	15	
			0.074	0.151	0.223	**0.231**	0.170	0.084	0.028	0.014

Notes. Accept-reject sampling scheme. Indicator matrices are sampled from their unrestricted prior distribution and only those satisfying the identification restrictions are kept. Simulations based on 10^7 draws. Acceptance rate in last column.

sampler (Algorithm 1) for 50,000 iterations, starting with the maximum number of potential factors and a random structure. The value of $\boldsymbol{\Delta}$ from the last iteration is then saved and used as a starting value, where only the identified factors (those with at least three dedicated measurements) are kept as active factors. The remaining measurements—those dedicated to unidentified factors—are assumed to be initially allocated to none of the identified factors.

Baseline Comparison to Classical EFA. We also perform classical exploratory factor analysis on the simulated data sets and compare the results to those obtained with BEFA. In a first step, we apply various criteria to select the number of factors. As explained in the next paragraph describing the results, no clear picture emerges and these criteria do not manage to uncover the dimension of the latent structure in a consistent way. Therefore, in a second step we run the factor analysis conditional on the true number of factors. Maximum likelihood factor analysis is implemented, as this classical factorization method is closest to our Bayesian approach.[31] The results are finally rotated using a Promax rotation, which generates a sparse factor loading matrix and is thus in line with our approach. Similarly to BEFA, a reordering of the columns has to be done to allow a comparison of the estimated factor loading matrix to the true one. This is done by first setting to zero all factor loadings

[31] Classical estimation was carried out with the R Statistical Package (R Core Team, 2013).

lower than 0.2 in magnitude, and then reordering the columns to match the true structure of the factor loading matrix as close as possible.

This comparison helps us assess the benefits of our approach over classical factor analytic methods. Nevertheless, the comparison should be done carefully, due to some differences in the implementation of the two approaches. Since classical criteria provide no conclusive answers to the selection of the number of factors (see Table 5), the maximum likelihood estimation presented in Table 3 is conducted conditional on the true number of factors. BEFA, on the contrary, estimates the number of factors using little prior information—the only prior information is conveyed by the prior distribution of the indicators, so as to generate plausible values for the number of factors (see Table 2).[32] The maximum likelihood approach does not explicitly use the information that the measurements are dedicated, contrary to BEFA. However, the cutoff value used to set the factor loadings to zero (0.2) in the classical approach is based on the minimum value the factor loadings can take in our data generating process. In real-data applications, practitioners do not have this information and would typically fix this cutoff at a higher value (e.g., 0.5), thus changing dramatically the final structure of the factor loading matrix. BEFA does not rely on such cutoff values and therefore does not make use of this information.

Monte Carlo Results. The results of the Monte Carlo experiments on our eight artificial models are summarized in Table 3. To grasp the performance of our MCMC sampler, we compute different statistics based on posterior modes and on the highest probability model (HPM), which corresponds to the indicator matrix most often visited by the sampler across MCMC iterations.

The BEFA algorithm manages to recover the true structure of the factor loading matrix in virtually all cases, as indicated by the hit rates that are all very high. The larger the model, the more difficult the factor search, especially in this context of very noisy data. More data available enables the sampler to better recover the full 0/1 pattern of the indicator matrix, as indicated by the larger hit rates for $N = 1000$ compared to $N = 500$ in the column $\boldsymbol{\Delta}^{\text{H}}$ for all models. Measurements that actually belong to the model are almost never wrongly discarded (first four models), and extra measurements—those that are uncorrelated with the other measurements—are retrieved very accurately (last four models). This last result is obtained thanks to the hierarchical prior on the indicator matrix with measurement-specific

[32]Nonetheless, we show in the Web Appendix that the impact of this prior distribution is negligible.

Table 3: Monte Carlo results from Bayesian EFA (unknown number of factors *a priori*, but no more than K) and from Classical EFA (maximum likelihood estimation with Promax rotation, conditional on true number of factors K_0) for models $\mathcal{M}(M, K_0, D, D_0)$ with $N = 500$ and 1000. 100 Monte Carlo replications.

Model	N	Bayesian EFA								Classical EFA (cond. on K_0)					
		Hit Rates					Averages		Hit Rates			Averages		
		K	$\widetilde{K}_1$	K_1^{H}	$\boldsymbol{\Delta}^{\text{H}}$	$\widetilde{D}_0$	D_0^{H}	K_1^{H}	D_0^{H}	$\boldsymbol{\Delta}$	$n_{\boldsymbol{\Delta}}$	D_0	$n_{\boldsymbol{\Delta}}$	$\overline{n}_{\boldsymbol{\Delta}}$	D_0
$\mathcal{M}(15,3,5,0)$	500	5	1.00	1.00	0.97	1.00	1.00	3.00	0.00	0.87	0.95	0.92	14.87	0.11	0.08
	1000	5	1.00	1.00	1.00	1.00	1.00	3.00	0.00	0.96	1.00	0.96	14.94	0.00	0.06
$\mathcal{M}(36,6,6,0)$	500	9	0.98	0.98	0.95	1.00	1.00	6.02	0.00	0.42	0.52	0.72	35.25	2.38	0.29
	1000	9	0.99	1.00	1.00	1.00	1.00	6.00	0.00	0.73	0.85	0.85	35.58	0.83	0.17
$\mathcal{M}(72,9,8,0)$	500	12	0.99	0.99	0.93	0.97	0.96	9.01	0.04	0.13	0.20	0.53	69.99	6.51	0.72
	1000	12	0.96	0.96	0.96	1.00	1.00	8.96	0.00	0.39	0.56	0.57	70.79	3.69	0.57
$\mathcal{M}(120,12,10,0)$	500	18	0.98	0.98	0.82	0.99	0.99	11.98	0.01	0.00	0.01	0.25	111.71	24.23	1.46
	1000	18	0.92	0.92	0.89	1.00	1.00	11.94	0.00	0.12	0.24	0.34	115.59	11.57	1.28
$\mathcal{M}(17,3,5,2)$	500	5	1.00	1.00	0.94	0.94	0.95	3.00	1.99	0.90	0.93	0.94	14.92	0.32	2.00
	1000	5	1.00	1.00	0.99	0.99	0.99	3.00	1.99	0.92	0.98	0.93	14.94	0.05	2.05
$\mathcal{M}(39,6,6,3)$	500	9	0.98	0.98	0.91	0.96	0.97	6.02	2.99	0.38	0.47	0.66	34.97	3.28	3.12
	1000	9	1.00	1.00	0.99	0.98	0.99	6.00	2.99	0.56	0.68	0.74	35.56	1.33	3.24
$\mathcal{M}(76,9,8,4)$	500	12	1.00	1.00	0.84	0.92	0.95	9.00	3.99	0.05	0.14	0.40	69.22	9.53	4.38
	1000	12	0.95	0.96	0.91	0.96	0.96	9.00	3.96	0.31	0.40	0.63	70.05	6.33	4.33
$\mathcal{M}(125,12,10,5)$	500	18	0.98	0.98	0.78	0.90	0.91	11.98	5.03	0.00	0.00	0.23	111.57	25.28	5.91
	1000	18	0.95	0.95	0.93	0.96	0.98	11.95	4.98	0.19	0.28	0.45	116.53	10.87	5.74

Notes. M is the total number of measurements, K_0 the true number of latent factors, D the number of measurements dedicated to each factor, D_0 the number of uncorrelated measurements, and N the sample size. Hit rates (i.e., proportion of Monte Carlo replications where the statistic is equal to the true value) and Monte Carlo averages are reported for different statistics: Number of factors recovered K_1, number of zero rows in the factor loading matrix D_0, full 0/1 pattern of the indicator matrix $\boldsymbol{\Delta}$. For Bayesian EFA, these statistics are computed based on the posterior mode ($\widetilde{K}_1$ and $\widetilde{D}_0$) and and on the highest probability model (K_1^{H}, $\boldsymbol{\Delta}^{\text{H}}$ and D_0^{H}). Monte Carlo averages over replications with a M-H rate larger than 0.8. For Classical EFA: Columns of factor loading matrix reordered to match true matrix, loadings lower than 0.2 in absolute are set to zero. $n_{\boldsymbol{\Delta}}$ is the number of non-zero indicators correctly recovered, and $\overline{n}_{\boldsymbol{\Delta}}$ the number of zero indicators wrongly recovered as non-zero.

parameters τ_{0m}, which introduces more flexibility in the estimation of the number of zero rows of the indicators matrix, especially in large models (see Subsection 2.3.1).[33]

Table 4 display some information assessing the numerical efficiency of our sampler. In most cases, Metropolis-Hastings acceptance rates are very high. Low acceptance rates indicate ill-convergence, as the sampler keeps proposing nonidentified models that never get accepted. In such cases, it is recommended to restart the sampling with different starting values. To gauge the numerical accuracy of our sampler, we compute inefficiency factors for the correlations of the factors, the top elements of the factor loading matrix, as well as the idiosyncratic variances corresponding to the highest posterior probability models (HPM). Each of these inefficiency factors is computed as the inverse of the relative numerical efficiency (Geweke, 1989), and measures the number of draws required to achieve the same numerical precision as an independent sample from the target distribution.[34] These factors are close to 1 in all cases, revealing a very good mixing of our sampler.[35] These good properties are obtained thanks to the marginal data augmentation sampling scheme used for the correlation matrix, and also to the intermediate steps in augmented models that are not saved for posterior inference and therefore induce a thinning of the Markov chain. Inference was conducted with a code written in Fortran for improved speed, and computational time was assessed on 2.66GHz Intel Xeon CPUs. Running times are displayed in minutes, and correspond to the total number of 40,000 MCMC iterations, each iteration being made of $2S$ intermediate steps in augmented models (10 intermediate steps on average).

The last columns of Table 3 show the results obtained from Maximum Likelihood estimation of the factor models with Promax rotation run on the same data sets. This approach is clearly outperformed by our BEFA method. It turns out to perform reasonably well on small models, but exhibits difficulties in recovering the true pattern of the indicator matrix when model size increases—although it is run conditional on the true number of latent factors and the true value of the minimum factor loading is used as threshold. The larger the model, the worse the performance: Too many correlated measurements turn out to be discarded (cf. columns for D_0) and some factor loadings equal to zero in the true model are estimated as different from zero (cf. column for $\overline{n}_{\boldsymbol{\Delta}}$). The comparison between the two approaches is

[33] We ran the same simulations with the initial prior specification on τ assuming a common parameter τ_0 across measurements, and as expected, the number of uncorrelated measurements D_0 was always underestimated, especially in large models. See Web Appendix for more details.

[34] For example $100,000$ draws from a sampler with an inefficiency factor of 10 will have the same numerical accuracy as $10,000$ draws from an independent sample. Inefficiency factors computed as explained in Kastner and Frühwirth-Schnatter (2014).

[35] Larger inefficiency factors would be obtained if they were not calculated for HPM—i.e., if they took into account model uncertainty due to the unknown structure of the factor loading matrix $\boldsymbol{\alpha}$. However, researchers are usually interested in the final structure of $\boldsymbol{\alpha}$ (HPM in this case), hence the results reported.

Table 4: Monte Carlo Experiment for BEFA: Sampling efficiency and computational time.

					Inefficiency factors			Time
Model	N	M-H acc.	Pr(acc $>$ .8)	p^{H}	$\boldsymbol{R}$	$\boldsymbol{\alpha}^{\text{lead}}$	$\boldsymbol{\Sigma}$	in min.
$\mathcal{M}(15,3,5,0)$	500	0.96	0.98	0.97	1.11	1.10	1.10	1.88
	1000	0.99	1.00	1.00	1.07	1.04	1.05	3.29
$\mathcal{M}(36,6,6,0)$	500	0.95	0.96	0.91	1.11	1.09	1.09	6.91
	1000	0.99	1.00	0.99	1.04	1.02	1.02	10.42
$\mathcal{M}(72,9,8,0)$	500	0.98	0.99	0.92	1.06	1.05	1.05	19.93
	1000	1.00	1.00	0.98	1.02	1.01	1.01	27.16
$\mathcal{M}(120,12,10,0)$	500	0.96	0.99	0.82	1.09	1.08	1.08	67.55
	1000	0.99	0.99	0.97	1.02	1.01	1.01	84.14
$\mathcal{M}(17,3,5,2)$	500	0.95	0.96	0.86	1.14	1.13	1.13	1.99
	1000	0.97	0.98	0.91	1.08	1.07	1.07	3.48
$\mathcal{M}(39,6,6,3)$	500	0.94	0.95	0.77	1.14	1.12	1.12	7.52
	1000	0.98	0.98	0.84	1.07	1.06	1.06	11.20
$\mathcal{M}(76,9,8,4)$	500	0.96	0.97	0.66	1.10	1.09	1.09	21.11
	1000	0.99	0.99	0.82	1.03	1.01	1.01	28.84
$\mathcal{M}(125,12,10,5)$	500	0.95	0.98	0.58	1.11	1.10	1.10	72.36
	1000	0.99	1.00	0.77	1.02	1.01	1.01	89.53

Notes. Metropolis-Hastings acceptance rate (M-H acc.), probability of highest probability model (p^{H}), proportion of Monte Carlo replications with a M-H rate larger than 0.8, inefficiency factors for the off-diagonal elements of the correlation matrix $\boldsymbol{R}$, for the leading elements of the factor loading matrix $\boldsymbol{\alpha}^{\text{lead}}$ and for the idiosyncratic variances $\boldsymbol{\Sigma}$. The inefficiency factors are computed as the averages of the median of the corresponding values over the Monte Carlo replications corresponding to the highest probability model. Average computational time in minutes, for a total of 40,000 MCMC iterations for each experiment. Monte Carlo averages over replications with a M-H rate larger than 0.8.

thus striking, especially given the fact that although BEFA is run without knowing the true number of factors *a priori*, contrary to classical EFA, it still manages to perform better in recovering the true latent structure.

Finally, Table 5 shows the results obtained by applying to the same simulated data methods routinely used in psychometrics and econometrics to select the number of components/factors.[36] While, as seen in Table 3, the BEFA algorithm displays remarkably high hit rates, the different classical criteria are not able to recover the dimensionality of the true latent structure in a consistent way. In particular, while most of the methods succeed in

[36]A brief description of the various classical methods used in this section for selecting the number of components/factors and for performing rotation is provided in the Web Appendix. The scree plots displaying the average eigenvalues across the 100 Monte Carlo replications for each model are also shown there.

recovering it for the simplest models with three factors, their performance varies between under- (in the case of the Velicer and of the Onatski method, and of the Bayesian Information Criterion) and over-extraction (in the case of the Kaiser criterion) for the higher-dimensional models. In general, doubling the number of observations from 500 to 1,000 allows a more accurate selection of the number of factors, while including in the data extra measurements uncorrelated with the others (as in the last four models) leads to an even greater degree of over-extraction.

We now apply our methodology to real data for the estimation of a high-dimensional factor model.

4.2 Empirical Analysis of the BCS Data

This section of the paper applies our method to data on cognitive, psychological and health measurements. Classical Exploratory Factor Analysis is widely used to boil down high dimensional data on psychological traits to interpretable scales. This is the method used to obtain the Big Five[37] (see Goldberg, 1990). We estimate the structure of cognitive, psychological and physical traits in childhood using the BEFA approach. We then show which alternative structures are obtained by the methods traditionally used.

Data. We apply our method to data from the British Cohort Study (BCS), which has been widely used in an interdisciplinary literature on the effects of early life conditions on adult outcomes. The BCS is a longitudinal survey following all babies born in a particular week of April 1970 in the United Kingdom. A wealth of information has been collected at multiple ages on the cohort members' cognitive, behavioral and physical development, their family and school environment, and their labor market and life outcomes. For this application, we use information on family background characteristics from the birth sweep, and on 131 cognitive, behavioral and health measurements—28 binary and 103 continuous—at age 10, to estimate the structure of childhood traits for the male cohort members.[38]

Prior Specification and MCMC Tuning. We run our algorithm on this data set and assume that the number of underlying factors does not exceed 20, (so $K = 20$).[39] We adopt a prior specification that is similar to the one used in the Monte Carlo study, assuming $a_m^0 = 0$ and $A_0 = 3$, and for the continuous measurements $c_0 = 2.5$ and C_m^0 specified as in

[37] In psychology, the Big Five personality traits are five broad domains or dimensions that are used to describe human personality, and that are based on the Five Factor Model (FFM)(Costa and McCrae, 1992). The Big Five are Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism (OCEAN).

[38] Full details on the data and the measures we use are in Appendix B.

[39] Since we find 13 factors, there is no need to rerun with a larger maximum number of factors.

Table 5: Monte Carlo results from Classical Methods to Select the Number of Components/Factors for models $\mathcal{M}(M, K_0, D, D_0)$ with N=500 and 1000. 100 Monte Carlo replications.

		Number of components						**Number of factors**									
		Velicer		Opt. Coord.		Kaiser		Opt. Coord.		Kaiser		Akaike IC		Bayesian IC		Onatski	
Model	N	Mode	HR	Mode	HR	Mode	HR	Mode	HR	Mode	HR	Mode	HR	Mode	HR	Mode	HR
$\mathcal{M}(15,3,5,0)$	500	3	0.65	3	0.92	3	0.93	2	0.28	3	0.96	3	0.82	3	0.95	3	0.52
	1000	3	0.63	3	0.90	3	0.96	2	0.23	3	0.96	3	0.79	3	0.99	3	0.65
$\mathcal{M}(36,6,6,0)$	500	5	0.25	7	0.29	7	0.20	4	0.01	6	0.77	6	0.88	5	0.25	5	0.05
	1000	5	0.24	6	0.63	6	0.64	4	0.01	6	0.75	6	0.86	6	0.57	5	0.22
$\mathcal{M}(72,9,8,0)$	500	7	0.18	13	0.12	15	0.00	7	0.00	12	0.00	9	0.77	7	0.00	5	0.03
	1000	8	0.09	11	0.13	12	0.03	6	0.00	9	0.79	9	0.91	8	0.60	5	0.09
$\mathcal{M}(120,12,10,0)$	500	10	0.06	14	0.04	26	0.00	10	0.00	22	0.00	11	0.39	7	0.00	10	0.01
	1000	11	0.05	17	0.08	22	0.00	9	0.01	15	0.00	12	0.78	9	0.01	10	0.02
$\mathcal{M}(17,3,5,2)$	500	3	0.65	3	0.41	5	0.00	2	0.29	3	0.96	3	0.84	3	0.93	3	0.51
	1000	3	0.71	3	0.58	4	0.00	2	0.31	3	0.99	3	0.79	3	0.99	3	0.72
$\mathcal{M}(39,6,6,3)$	500	5	0.19	9	0.19	9	0.00	4	0.00	7	0.42	6	0.81	5	0.18	5	0.09
	1000	5	0.24	8	0.21	9	0.00	4	0.02	6	0.82	6	0.84	6	0.52	5	0.22
$\mathcal{M}(76,9,8,4)$	500	7	0.03	15	0.07	18	0.00	7	0.00	14	0.00	9	0.75	6	0.00	5	0.02
	1000	8	0.08	13	0.08	15	0.00	6	0.00	9	0.48	9	0.90	7	0.05	5	0.07
$\mathcal{M}(125,12,10,5)$	500	10	0.05	15	0.06	30	0.00	10	0.06	27	0.00	11	0.37	7	0.00	10	0.03
	1000	10	0.07	14	0.06	24	0.00	9	0.00	18	0.00	12	0.86	9	0.00	10	0.00

Notes. M is the total number of measurements, K_0 the true number of latent factors, D the number of measurements dedicated to each factor, D_0 the number of uncorrelated measurements, and N the sample size. Opt. Coord. = Optimal Coordinates. IC = Information Criterion. HR = Hit Rate. We use the eigenvalues of the raw correlation matrix to find the number of components (when applying the Velicer, Optimal Coordinates and Kaiser methods), and the eigenvalues of the reduced correlation matrix to find the number of factors (when applying the Optimal Coordinates and Kaiser methods). To construct the reduced correlation matrix, we use the squared multiple correlations as estimates of the communalities. The Akaike and Bayesian Information Criteria are computed after having performed maximum likelihood factor analysis. For the Onatski method, we specify $k_0 = 1$ and $k_1 = 5$ for $\mathcal{M}(15,3,5,0)$ and $\mathcal{M}(17,3,5,2)$; $k_0 = 5$ and $k_1 = 10$ for $\mathcal{M}(36,6,6,0)$, $\mathcal{M}(39,6,6,3)$, $\mathcal{M}(72,9,8,0)$ and $\mathcal{M}(76,9,8,4)$; $k_0 = 10$ and $k_1 = 15$ for $\mathcal{M}(120,12,10,0)$ and $\mathcal{M}(125,12,10,5)$. For each test, the first column (Mode) shows the modal number of factors obtained across the 100 Monte Carlo replications; the second column (Hit Rate) shows the proportion of Monte Carlo replications where the number of factors is equal to the true value.

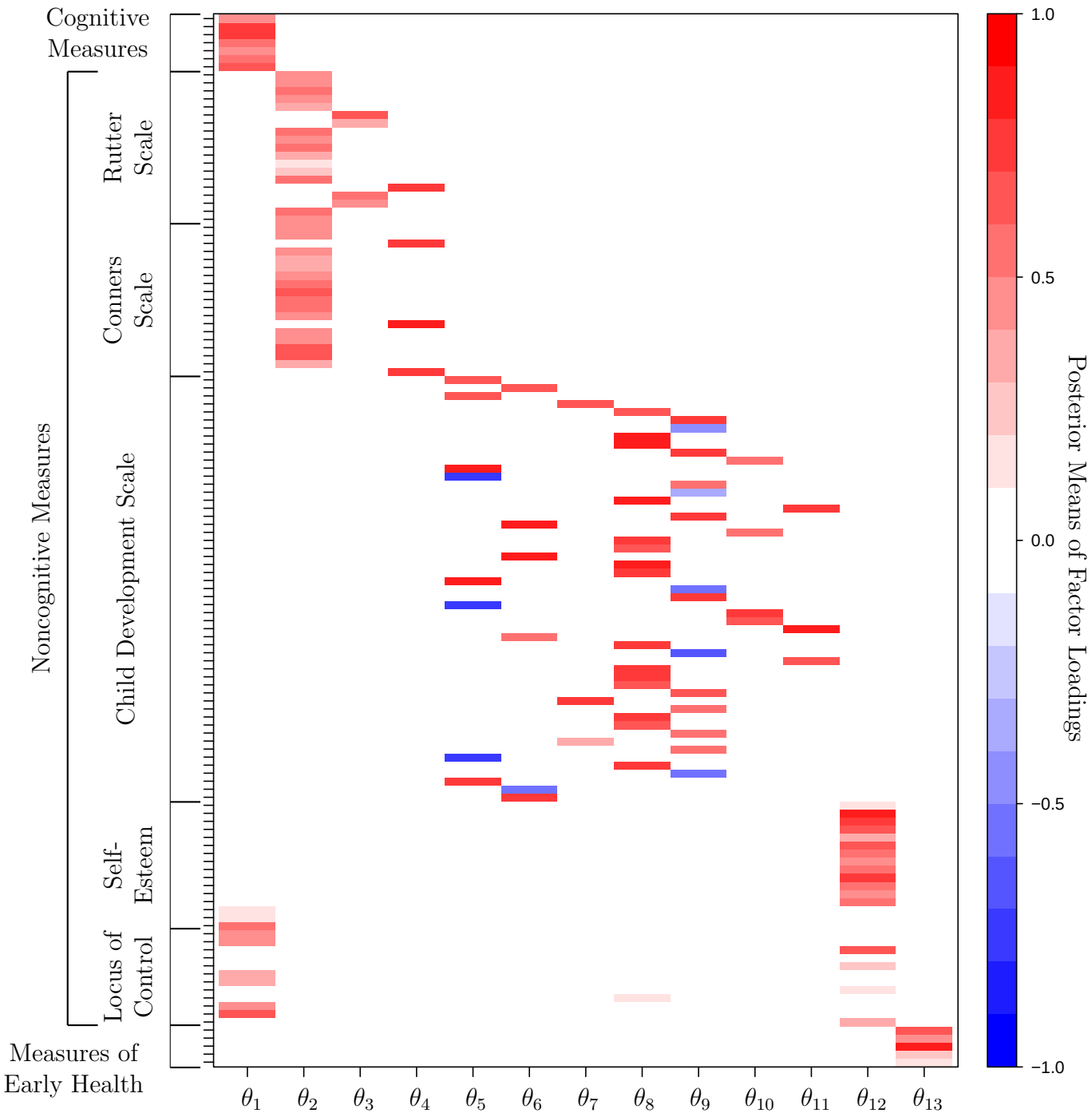
equation (17). The only differences worth pointing out are for the regression coefficients, the correlation matrix of the factors, and the indicator probabilities. We introduce covariates in our factor model to control for observed heterogeneity, and assume that the corresponding regression coefficients are *a priori* centered ($b_m^0 = \mathbf{0}$) with prior variance $\boldsymbol{B}_m^0 = 3\boldsymbol{I}$.[40] To hinder factor splitting, which happens to be a problem in our application when assuming a uniform prior on the individual factor correlations, we increase the number of degrees of freedom to $\nu = K+5$. As shown in Figure 1, this value of ν shifts the prior distribution of the maximum correlation away from 1. The scale matrix $\boldsymbol{S}$ is specified as stochastic to implement the Huang and Wand (2013) prior, and its diagonal elements are allowed to take relatively large values to enhance mixing by fixing $A_k = 100$. Finally, the prior on the indicator weights is specified with measurement-specific parameters τ_{0m} assumed to have a symmetric prior Beta distribution ($\kappa_0 = \xi_0 = 0.1$), and for the included measurements the Dirichlet prior is specified with concentration parameter $\kappa = 0.5$, a prior similar to the one used for the largest model with 125 measurements in our Monte Carlo study (see Table 2). We start the algorithm with a single factor and run the sampler for $100,000$ iterations, where only the last $40,000$ ones are used for posterior inference.[41] For the factor selection, $2S$ intermediate steps are performed at each MCMC iteration, with $S = 1 + \phi$ and $\phi \sim \mathcal{P}oisson\,(4)$. We run the MCMC sampler several times with different starting values to check it converges to the same solution. After sampling, the MCMC draws are post-processed following the strategy described in Subsection 3.4 to solve the sign and column switching problems and make interpretation possible.

Empirical Results. The main results are presented in Figure 2, which displays the posterior means of the factor loadings in the highest probability model (HPM)—the model that corresponds to the indicator matrix $\boldsymbol{\Delta}$ that is visited most often by the algorithm. In our application, the posterior probability of the HPM is 0.42. These results show that the method succeeds in condensing the information contained in the data in a concise and interpretable way. BEFA uncovers 13 factors (out of an admissible set of $K = 20$) from the 131 measurements recorded from multiple sources on the development of the child at age 10. The factor loading matrix should be interpreted jointly with Figure 3 that shows the posterior correlations among the estimated factors and gives more insights into the interrelations between the latent constructs.

[40] See Appendix B for scaling of the covariates.

[41] We resort to a long burn-in period for the empirical application, as the pre-MCMC stage based on the unrestricted algorithm turned out to produce too many nonidentified factors that could not be used to generate a sensible starting value for $\boldsymbol{\Delta}$.

Figure 2: BEFA, Posterior Factor Loading Matrix in the BCS.



The factors capture the following traits (interpretation done *a posteriori*):

θ_1	Cognitive Ability,	θ_2	Behavioral Problems [M],	θ_3	Anxiety [M],
θ_4	Hyperactivity [M],	θ_5	Attention Problems [T],	θ_6	Anxiety [T],
θ_7	School Phobia [T],	θ_8	Conduct Problems [T],	θ_9	Motor Coordination Prob. [T],
θ_{10}	Depression [T],	θ_{11}	Concentration Prob. [T],	θ_{12}	Positive Sense of Self [C],
θ_{13}	Body Build.				

Notes. The 131 measurements (tick marks on the vertical axis) are in the order specified in Appendix B. [M] refers to traits extracted from items evaluated by the mother, [T] by the teacher, [C] by the child. Active factors only are displayed, out of a maximum of 20 potential factors.

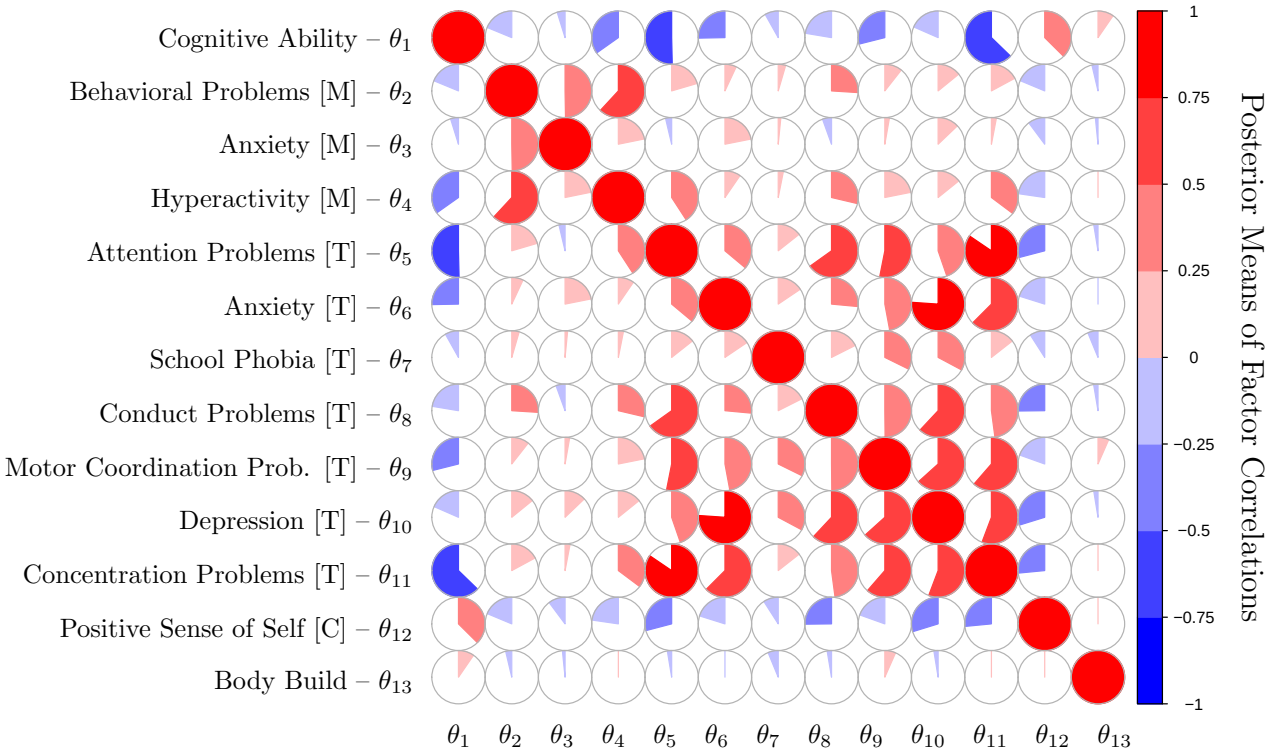
First of all, the measurements are clearly allocated to one of three broad categories—cognitive, noncognitive and health. All intelligence test scores load on a single factor, which we term cognitive ability (θ_1);[42] likewise, all physical measurements load on a separate factor, hence named Body Build (θ_{13}). Most importantly, the numerous measures belonging to the five noncognitive scales (the Rutter, Conners, Child Developmental, Self-Esteem and Locus of Control scales) are allocated to 11 different factors in such a way that items describing the same trait consistently load on the same factor. In this way, each factor beyond the first one can be clearly named as a child mental health problem or facet of temperament, as shown in the columns of Figure 2 (θ_2–θ_{12}). For example, the factor we call "Attention Problems [T]" (θ_5) is loaded by all teacher-reported items denoting inability to pay attention in class. Second, measurements collected from different subjects (mothers, teachers, and the children themselves) load on separate factors, although some of them use exactly the same wording.[43] Third, Figure 3 shows that the estimated correlations among the factors are informative:[44] in addition to the two main clusters of inter-correlated mother-(θ_2–θ_4) and teacher- (θ_5–θ_{11}) reported traits, BEFA also succeeds in uncovering meaningful correlations across traits derived from reports by different informants. For example, the correlation between cognitive ability (θ_1), as measured by intelligence test scores administered to the child, and attention problems (θ_5), as measured by teacher-reported items, is -0.504. And the low correlation between mother- and teacher-reported traits is also consistent with a consolidated literature in child psychology, starting from the seminal study of Achenbach et al. (1987), who report that correlations of ratings are low between informants who play different roles with children.

[42]Items from the locus of control scale also load on this factor. While this might seem *prima facie* unusual, it is not actually uncommon in the literature. Costa and McCrae noticed "Many lexical studies show that some aspects of rated or self-reported intelligence (e.g., logical, foresighted vs thoughtless, imperceptive) also load on a Conscientiousness factor; we view these as reflections of Competence. We would also hypothesize that locus of control would be related to this facet." (Costa et al., 1991). Additionally, also Van Lieshout and Haselager (1994) and Mervielde et al. (1995) obtain childhood factors loading on both intellectual capacity/intelligence and Conscientiousness. Finally, von Stumm et al. (2009), analysing these same data, also notice a substantial overlap of locus of control and intelligence. They hypothesize this may be partially due to the shared cognitive-based setting of assessment (i.e., in school under teacher's supervision). Alternatively, like Costa et al. (1991), they speculate that these scales may tap into the same dimension of individual differences. Intelligence enables learning and understanding, which facilitate pupils' school performance and academic achievement. This encourages a sense of personal competency and, thus, students are likely to attribute school achievements to their own ability and effort rather than external circumstances. In our results, all the locus of control items which load on factor 1 specifically refer to academic performance, attesting that the measurement of locus of control in the BCS 1970 is closely linked to school experiences.

[43]This occurs in the case of the Child Developmental Scale, which was specifically developed for inclusion in the BCS by selecting appropriate items from the Rutter and Conners instruments, and adding a few additional ones—such as motor coordination problems—to make the scale a more comprehensive measure of child development. The list of items with the same wording and the different factors they load on is shown in the Web Appendix. The detailed description of each item by which each factor is loaded is also reported there.

[44]Posterior standard errors for the estimated correlations are displayed in Table A3.4 in the Web Appendix.

Figure 3: BEFA: Posterior Correlation Matrix of the Factors in the BCS Application.



Notes. Each pie represents the correlation between the corresponding factors, clockwise for positive values and counterclockwise for negative values. [M] refers to traits extracted from items evaluated by the mother, [T] by the teacher, [C] by the child. Active factors only are displayed, out of a maximum of 20 potential factors. For standard errors and credible intervals, see Web Appendix.

Comparison with Estimates from Classical Methods. We now compare the performance of our procedure with that of approaches traditionally used in the applied literature.[45] Given the lack of a commonly accepted method of aggregation, different studies summarize the available information in many different ways, and often arrive at different conclusions, even when analyzing the same data. First and foremost, all studies make *a priori* judgments on which sets of scales to aggregate: no previous study has analyzed all the information available in the data as we do here. At the initial stage, researchers usually define broad categories—such as cognition, personality and health—then eventually define sub-categories (e.g., verbal and mathematical intelligence, conduct or attention problems). This approach may be appropriate when *a priori* information is available to the researcher. Then, analysts use their method of choice to condense the information available within each of these pre-defined categories. The two most commonly used approaches are: (1) construction of simple

[45] A brief description of the various classical methods used in this section for selecting the number of components/factors and for performing rotation is provided in the Web Appendix.

sums or averages; (2) Exploratory Factor Analysis (EFA), with the extraction of principal components or factors.

A first common approach to aggregation is to take sums or unweighted averages, either of different scales belonging to a broad category (e.g., all cognitive scales), or of different items belonging to the same scale (e.g., all items belonging to the self-esteem scale), as done in Murasko (2007), Gale et al. (2008) and Kaestner (2009), among others. This simple procedure makes two strong assumptions: equal weighting of items (i.e., all measures are assumed to incorporate the same share of information about the latent factors), and absence of measurement error. Both of these assumptions are at odds with the data.[46] On the one hand, different measurements associated with the same factor clearly have different factor loadings (Figure 2). On the other hand, we find substantial measurement error in the measurements (Figure 4). This provides evidence that, at least when using the BCS data, unweighted aggregates are not an adequate representation of the latent structure of childhood traits.

Another approach commonly adopted is to extract principal components or factors. Although the two methods are conceptually different, they are often used interchangeably in the applied literature, when there is need for dimensionality reduction. For example, Feinstein (2000), Blanden et al. (2007), Gale et al. (2009), Jones et al. (2011) and Dohmen et al. (2012) all extract principal components, while von Stumm et al. (2009), Baron and Cobb-Clark (2010), Antecol and Cobb-Clark (2010), Helmers and Patnam (2011) and Fiorini and Keane (2012) extract factors, although they deal with similar applications and sometimes even use the same data.

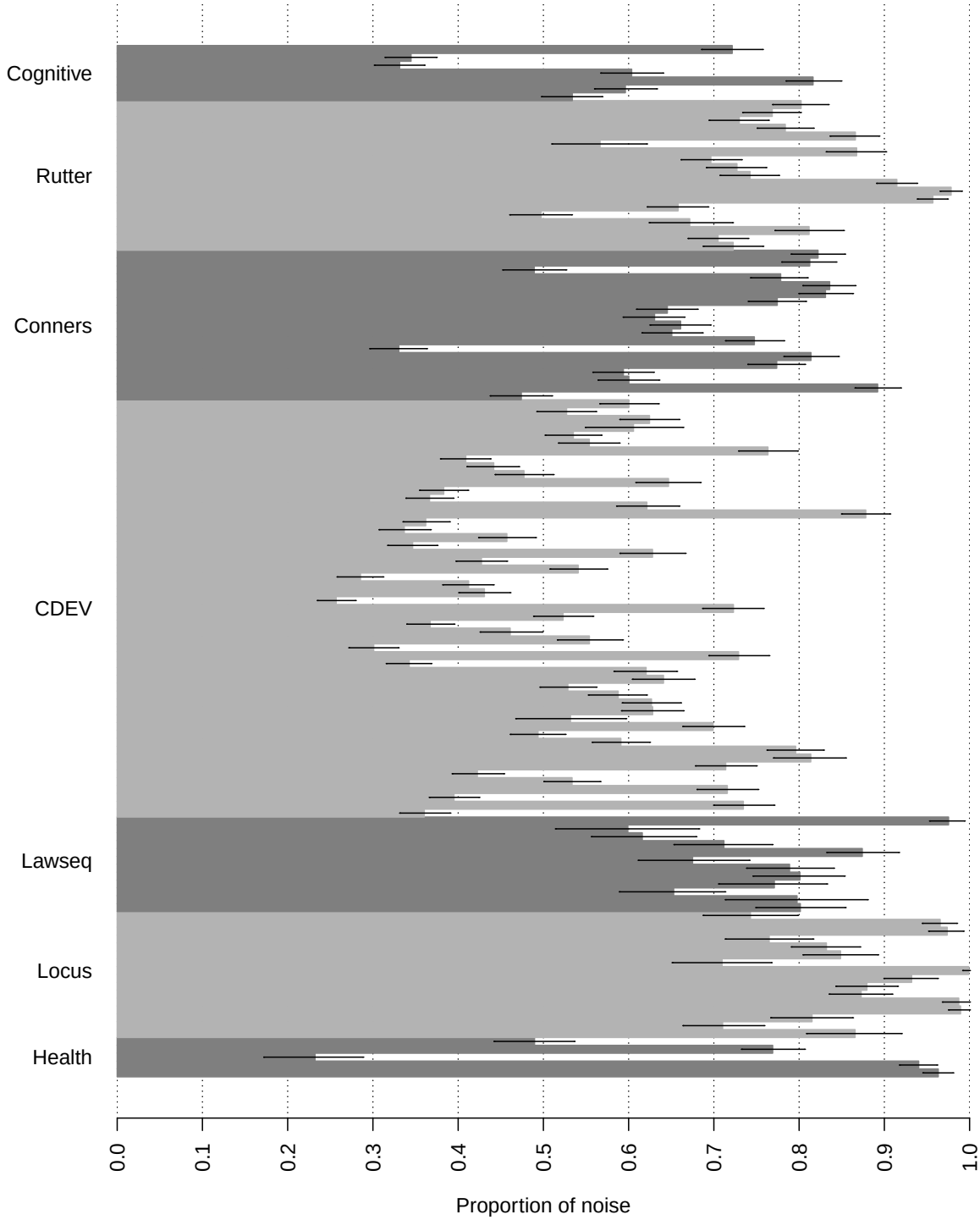
When components are extracted, error in the measurements is not accounted for. When extracting factors, instead, the analyst only analyzes the variability in the observed measurements which can be explained by the common factors not affected by measurement error. Stated differently, components extraction is based on an eigenvalue decomposition of the raw correlation matrix (Jolliffe, 1986), while factor extraction is applied on the "reduced" correlation matrix with measurement error variance removed (the one based on the factor covariance).[47]

It is instructive to compare the steps involved across the various methods. While BEFA is a unified single step procedure, traditional approaches involve *multiple* stages: a first step in which the number of components/factors is selected, a second step in which components/factors are extracted (conditional on the number selected in the first step), and a third

[46] See Cunha and Heckman (2008) for an exploration of these issues.

[47] In practice, the two methods will yield similar results when the values of the communalities are relatively high (Fabrigar et al., 1999).

Figure 4: BEFA, Proportion of total variance of measurements due to noise.



Notes. Segments on top of bars represent the corresponding 95% highest posterior density intervals. Measurements are ordered as follows, from top to bottom: Cognitive items (PLCT, FMT, SERT, BASTM, BASTRD, BASTS, BASTWD), Rutter1 to Rutter 19, Conners1 to Conners19, Child development scale (CDEV1 to CDEV53), Self-esteem (Lawseq1 to Lawseq12), Locus of Control (Locus1 to Locus16) and Health (Height, Head, Weight, Bpsys, Bpdias). See Appendix B for details.

step in which rotation is performed to search for a simple structure.[48] Arbitrary decision rules are involved at each step. Several criteria are available to select the dimensionality of the latent structure, to extract the factors (Gorsuch, 1983), and to rotate the resulting loading matrices (Jennrich, 2001, 2002). If a simple structure does not emerge in a first round, classical Exploratory Factor Analysis procedures also involve further steps, in which measurements weakly loading on factors are iteratively eliminated on the basis of arbitrary threshold rules, until a stable solution with only single loaders is achieved. The elimination criterion is also usually based on the magnitude of the loadings, without accounting for their statistical significance.

BEFA performs all of these steps in one coherent Bayesian procedure, where the dimension of the latent structure is estimated jointly with the allocation of the measurements to the factors. This is in contrast with traditional approaches in which the various steps are performed sequentially, and each of them requires *ad hoc* judgments, which affect the final outcome, as shown in Table 7.[49]

First, as already seen in the previous section with the application to the simulated data, the choice of the method used to select the dimensionality of the latent structure is not innocuous. Table 6 shows that the number of components/factors estimated from the raw measurements ranges from a minimum of 6 when using the Scree plot and the Onatski (2009) method, to a maximum of 72 when applying the Akaike Information Criterion.[50] It also shows that each method selects a number of components bigger than the number of factors. Because component extraction does not discriminate between common and unique variance, spurious components/factors are likely to be extracted. Additionally, using raw or residualized measurements[51] also makes a difference, since in the latter case a smaller number of components/factors is usually selected. This might suggest that, when using raw measurements, spurious components are likely to be extracted.

[48]The procedure of rotation identifies blocks of measures that within blocks are strongly correlated with one component/factor (i.e., satisfy *convergent validity*) but are weakly correlated with other components/factors across blocks (i.e., satisfy *discriminant validity*).

[49]Discarding measurements is an intuitively unsatisfactory procedure but it is an essential part of Exploratory Factor Analysis. (See, e.g., Gorsuch, 2003). At the same time, the procedure used in this paper can be faulted by assuming that each measurement loads on at most one factor. In future work, we plan to relax this requirement.

[50]Scree plots of the eigenvalues from both the raw and the reduced polyserial correlation matrix are shown in the Web Appendix. It is evident that, in both cases, no clear separation or "elbow" emerges.

[51]We define residualized measurements as the residuals of a linear regression of the measurements on the seven covariates ($\boldsymbol{X}$) which are included in the BEFA measurement system. We use a linear probability model for the binary measurements. The covariates included are mother's age at birth, mother's education at birth, father's high social class at birth, total gross family income at age 10, an indicator for broken family, the number of previous livebirths, and the number of children in the family at age 10. More details are provided in Appendix B.2.

Table 6: Classical Methods to Select the Number of Components/Factors

Method	**Number of components**		**Number of factors**	
	Raw Measurements	Residualized Measurements	Raw Measurements	Residualized Measurements
Cattell's Scree Plot	6	4	6	4
Onatski	n/a	n/a	6	5
Velicer's Rule	12	11	n/a	n/a
Optimal Coordinates	15	11	13	11
Kaiser's Rule	28	29	25	25
Akaike IC	n/a	n/a	72	47
Bayesian IC	n/a	n/a	21	18

Notes. IC = Information Criterion. We use the eigenvalues of the raw correlation matrix to find the number of components (when applying the Scree, Velicer, Optimal Coordinates and Kaiser methods), and the eigenvalues of the reduced correlation matrix to find the number of factors (when applying the Scree, Optimal Coordinates and Kaiser methods). To construct the reduced correlation matrix, we use the squared multiple correlations as estimates of the communalities. The Akaike and Bayesian Information Criteria are computed after having performed maximum likelihood factor analysis. For the Onatski method, we specify $k_0 = 3$ and $k_1 = 10$. We define residualized measurements as the residuals of a linear regression of the measurements on the covariates ($\boldsymbol{X}$) which are included in the BEFA measurement system (see Appendix B.2). We use a linear probability model for the binary measurements.

Second, in the classical approach, not only the criterion to detect the number of components/factors plays a role, but also the extraction and rotation methods have a non-negligible influence on the final structure. This is clearly visible in Table 7. Here we show the final number of components/factors and measurements which result by applying to the initial set of 131 measurements (both raw and residualized) different dimensionality selection criteria, extraction and rotation methods, and use the same set of rules to discard items, as suggested in Costello and Osborne (2005), and applied in Heckman et al. (2013).[52] The extraction methods applied are those of principal components (routinely used to construct ability measures),[53] principal factors (Gorsuch, 1983, 2003), applied among others in Antecol and Cobb-Clark (2010) and von Stumm et al. (2009), and maximum likelihood factor analysis (the method closer to BEFA). We then use two commonly applied methods of oblique rotation—promax and quartimin—that penalize departures from Thurstone's simple structure, and allow for correlated factors which are also accommodated in BEFA. Hence, for each set of measurements (raw or residualized), initial number of components/factors, extraction and rotation method, we apply the following rules. First, we exclude items with loadings

[52]Similar threshold rules to discard weakly-loading items and to interpret the resulting structure are applied by von Stumm et al. (2009) and Fiorini and Keane (2012).

[53]We use the component loadings, i.e., the eigenvectors scaled by the square root of the eigenvalues.

smaller than 0.5 in absolute value (to avoid the *weak-loading problem*), and also items with a loading of 0.32 or higher in absolute value (as suggested in Tabachnick and Fidell, 2001) on two or more factors (to avoid the *crossloading problem*). Second, we also exclude measurements in cases where only two of them load on a single factor (to avoid weakly-identified constructs). This restriction serves the same purpose as our identifiability condition (that at least three measurements must be dedicated to each factor). In the classical setup this condition is applied ex-post and in a sequence of steps subject to arbitrary choices, while in BEFA it is explicitly incorporated into the MCMC sampling scheme. This iterative procedure of components/factors selection, extraction, rotation, and elimination of measurements is repeated until no further items are dropped.

It is clear from Table 7 that both the choice of the initial number of components/factors to extract and the extraction/rotation method adopted have a substantial impact on the final structure achieved, when performing this iterative sequential elimination procedure.[54] The final structure achieved depends on both the chosen initial number of components/factors, and on the choices made at the various steps. It ranges between a minimum of 4 final factors and 34 measurements, to a maximum of 11 final factors and 76 measurements. Starting by selecting a smaller number of factors in general leads to retaining a smaller number of measurements. The choice itself among the different final configurations is not innocuous. While more parsimonious, a lower-dimensional structure would not display the richness of the childhood traits as uncovered by the BEFA approach. In almost all final configurations obtained with this procedure, the health factor has been discarded (with the exception of the structure obtained when extracting principal components with an initial number of 12); additionally, when using maximum likelihood factor analysis with 6 initial factors, the cognitive factor is discarded.

In sum, alternative classical approaches to aggregating high-dimensional data often make assumptions that are not supported by the data (equal weighting of items and absence of measurement error), or that lead analysts to discard many measurements loading on multiple factors. The multistage procedure of classical EFA is based on separate stage-wise choices of significance levels, criteria for selection of the dimension of the model, criteria for allocation of measurements to factors and criteria for estimating factor loadings made by the analysts at various steps of the process. Although the BEFA method requires some a priori judgments, it is a unified procedure that simultaneously chooses the dimension of the model, the allocation of measurements to factors and factor loadings using the same algorithm and tuning parameters.

[54] Browne (2001) was the first to show how different rotation criteria can influence factor pattern matrices.

Table 7: Final Number of Components/Factors (K^F) and of Measurements (M^F) Obtained by Applying Different Dimensionality Selection, Extraction and Rotation Methods and One Sequential Item Elimination Algorithm to the Initial Set of 131 Measurements.

Method	**Raw Measurements**					
Selection	Onatski (K^S =6)				Scree (K^S =6)	
Extraction	Maximum Likelihood		Principal Factors		Principal Components	
Rotation	Promax	Quartimin	Promax	Quartimin	Promax	Quartimin
K^F	6	6	6	6	6	6
M^F	46	46	41	40	56	56
Selection	Optimal Coordinates (K^S =13)				Velicer (K^S =12)	
Extraction	Maximum Likelihood		Principal Factors		Principal Components	
Rotation	Promax	Quartimin	Promax	Quartimin	Promax	Quartimin
K^F	10	10	10	10	11	11
M^F	64	56	73	66	74	76
Method	**Residualized Measurements**					
Selection	Onatski (K^S =5)				Scree (K^S =4)	
Extraction	Maximum Likelihood		Principal Factors		Principal Components	
Rotation	Promax	Quartimin	Promax	Quartimin	Promax	Quartimin
K^F	5	5	4	5	4	4
M^F	40	40	30	36	46	46
Selection	Optimal Coordinates (K^S =11)				Velicer (K^S =11)	
Extraction	Maximum Likelihood		Principal Factors		Principal Components	
Rotation	Promax	Quartimin	Promax	Quartimin	Promax	Quartimin
K^F	8	9	8	8	9	9
M^F	54	54	59	57	64	65

Notes. K^S is the number of components/factors selected at the beginning of the sequential item elimination procedure, according to the various methods as shown in Table 6. K^F is the final number of components/factors left at the end of the sequential item elimination procedure. M^F is the corresponding final number of measurements, out of an initial set of 131 measurements. For each set of measurements (raw or residualized), initial number of components/factors, extraction and rotation method, we apply the following rules. First, we exclude items with loadings smaller than 0.5 in absolute value (to avoid the *weak-loading problem*), and also items with a loading of 0.32 or higher in absolute value (as suggested in Tabachnick and Fidell, 2001) on two or more factors (to avoid the *crossloading problem*). Second, we also exclude measurements in cases where only two of them load on a single factor (to avoid weakly-identified constructs). All the final resulting structures are shown in the Web Appendix.

5 Conclusion

This paper develops and applies a new method—Bayesian Exploratory Factor Analysis (BEFA)—to constructing maximum posterior probability indices that summarize high-dimensional data by a low dimensional set of interpretable aggregates. We develop an integrated Bayesian framework to factor selection that simultaneously tackles several steps in building a factor model that are usually done sequentially: the choice of the dimension of the latent structure, the allocation of the measurements to the factors, as well as the estimation of the corresponding factor loadings.

Our method advances the traditional literature on Exploratory Factor Analysis. BEFA constitutes a significant departure from traditional factor-analytic methods by overcoming the intrinsic arbitrariness of the choices made by analysts in the various steps—from the choice of dimension to the extraction and rotation method. Ours is a coherent estimation framework. It is the first paper in the Bayesian literature to estimate a dedicated factor model with correlated factors, where the dimension of the factor structure is *a priori* unknown. Importantly, it links the two literatures, by invoking classical criteria to achieve identification, and by imposing identifying restrictions as an integral feature of the estimation algorithm.

We make several contributions in implementing our algorithm. To explore the parameter space, our sampler is allowed to navigate through expanded models where the identifying restrictions are relaxed. However, these intermediate steps are not used for posterior inference. They only serve as a computational tool. Eventually the algorithm only samples identified models. To draw the factors and their correlation matrix, marginal data augmentation methods as well as block sampling of the active and inactive factors have been adapted to our problem, to make it possible to sample these parameters and latent variables in a dimension-varying model.

We check the performance of our approach by a Monte Carlo experiment, and we show that it outperforms classical methods both for dimensionality selection, and identification of the true latent structure. Its applicability is demonstrated with an empirical study. We estimate the structure of the childhood mental and physical traits, and show that the method succeeds in producing interpretable aggregates. We compare its performance with that of several existing classical Exploratory Factor Analysis approaches. We show that application of classical Exploratory Factor Analysis can lead to different conclusions, depending on the choices made by the analysts at various steps of the process and the sequential item elimination rules used to achieve interpretability of the structure. Our method is a coherent, theoretically-based alternative.

Classical EFA discards data that load on multiple factors. Our version of BEFA does not discard data, except for measurements that do not load on any factor. However, the analysis of this paper assigns measurements to at most one factor. In research underway, we extend our approach to allow measurements to be allocated to multiple factors. This changes the identification and computation algorithm substantially and warrants a separate analysis.

A Details on MCMC Sampling

A.1 Proof of the Detailed Balance Condition for the MH Sampler with Intermediate Steps

To prove that the Markov chain resulting from the sampling scheme introduced in Section 3.1.2 leaves the distribution of $\boldsymbol{\vartheta}$ invariant, it is enough to show that the detailed balance condition holds for accepted moves. The probability of starting from a set of parameters $\widehat{\boldsymbol{\vartheta}}_0$ belonging to the identified set (i.e., $\widehat{\boldsymbol{\Delta}}_0 \in \mathcal{D}$), going through the sequence of intermediate states $\widehat{\boldsymbol{\vartheta}}_1, \dots, \widehat{\boldsymbol{\vartheta}}_{S-1}, \overline{\boldsymbol{\vartheta}}_S, \check{\boldsymbol{\vartheta}}_{S-1}, \dots, \check{\boldsymbol{\vartheta}}_0$, and finally accepting the final state $\check{\boldsymbol{\vartheta}}_0$ (i.e., if $\check{\boldsymbol{\Delta}}_0 \in \mathcal{D}$), can be shown to be the same as the probability of starting from the same state $\check{\boldsymbol{\vartheta}}_0$ (assuming it belongs to the identified set), moving to $\widehat{\boldsymbol{\vartheta}}_0$ through the same sequence of transitions, but *in reverse order*, and accepting $\widehat{\boldsymbol{\vartheta}}_0$ as the new identified state:

$$
\begin{aligned}
& p(\widehat{\boldsymbol{\vartheta}}_0) T_u(\widehat{\boldsymbol{\vartheta}}_0, \widehat{\boldsymbol{\vartheta}}_1) \dots T_u(\widehat{\boldsymbol{\vartheta}}_{S-1}, \overline{\boldsymbol{\vartheta}}_S) T_u(\overline{\boldsymbol{\vartheta}}_S, \check{\boldsymbol{\vartheta}}_{S-1}) \dots T_u(\check{\boldsymbol{\vartheta}}_1, \check{\boldsymbol{\vartheta}}_0)\, \delta_{\mathcal{D}}(\check{\boldsymbol{\Delta}}_0) && \text{(A1)} \\
& \quad = p(\widehat{\boldsymbol{\vartheta}}_0) \prod_{s=1}^{S} T_u(\widehat{\boldsymbol{\vartheta}}_{s-1}, \widehat{\boldsymbol{\vartheta}}_s) \prod_{s=1}^{S} T_u(\check{\boldsymbol{\vartheta}}_s, \check{\boldsymbol{\vartheta}}_{s-1})\, \delta_{\mathcal{D}}(\check{\boldsymbol{\Delta}}_0), \\
& \quad = \delta_{\mathcal{D}}(\widehat{\boldsymbol{\Delta}}_0) p_u(\widehat{\boldsymbol{\vartheta}}_0) \prod_{s=1}^{S} \frac{p_u(\widehat{\boldsymbol{\vartheta}}_s)}{p_u(\widehat{\boldsymbol{\vartheta}}_{s-1})} T_u(\widehat{\boldsymbol{\vartheta}}_s, \widehat{\boldsymbol{\vartheta}}_{s-1}) \prod_{s=1}^{S} \frac{p_u(\check{\boldsymbol{\vartheta}}_{s-1})}{p_u(\check{\boldsymbol{\vartheta}}_s)} T_u(\check{\boldsymbol{\vartheta}}_{s-1}, \check{\boldsymbol{\vartheta}}_s)\, \delta_{\mathcal{D}}(\check{\boldsymbol{\Delta}}_0), && \text{(A2)} \\
& \quad = \delta_{\mathcal{D}}(\widehat{\boldsymbol{\Delta}}_0) \prod_{s=1}^{S} T_u(\widehat{\boldsymbol{\vartheta}}_s, \widehat{\boldsymbol{\vartheta}}_{s-1}) \prod_{s=1}^{S} T_u(\check{\boldsymbol{\vartheta}}_{s-1}, \check{\boldsymbol{\vartheta}}_s)\, p_u(\check{\boldsymbol{\vartheta}}_0) \delta_{\mathcal{D}}(\check{\boldsymbol{\Delta}}_0), \\
& \quad = p(\check{\boldsymbol{\vartheta}}_0) T_u(\check{\boldsymbol{\vartheta}}_0, \check{\boldsymbol{\vartheta}}_1) \dots T_u(\check{\boldsymbol{\vartheta}}_{S-1}, \overline{\boldsymbol{\vartheta}}_S) T_u(\overline{\boldsymbol{\vartheta}}_S, \widehat{\boldsymbol{\vartheta}}_{S-1}) \dots T_u(\widehat{\boldsymbol{\vartheta}}_1, \widehat{\boldsymbol{\vartheta}}_0)\, \delta_{\mathcal{D}}(\widehat{\boldsymbol{\Delta}}_0), && \text{(A3)}
\end{aligned}
$$

where equation (A2) follows from the mutual reversibility condition of equation (23). Furthermore, both equation (A2) and equation (A3) use the fact that $p(\boldsymbol{\vartheta}) \propto p_u(\boldsymbol{\vartheta})\, \delta_{\mathcal{D}}(\boldsymbol{\Delta})$, see equation (24).

The detailed balance condition of the unrestricted MCMC move through the intermediate steps follows by integrating out the intermediate states $\widehat{\boldsymbol{\vartheta}}_1, \dots, \widehat{\boldsymbol{\vartheta}}_{S-1}, \overline{\boldsymbol{\vartheta}}_S, \check{\boldsymbol{\vartheta}}_{S-1}, \dots, \check{\boldsymbol{\vartheta}}_1$ on both sides of equation (A1), to provide the kernel of the transition from $\widehat{\boldsymbol{\vartheta}}_0$ to $\check{\boldsymbol{\vartheta}}_0$:

$$
\begin{aligned}
T_u(\widehat{\boldsymbol{\vartheta}}_0, \check{\boldsymbol{\vartheta}}_0) = \iint \dots \int & T_u(\widehat{\boldsymbol{\vartheta}}_0, \widehat{\boldsymbol{\vartheta}}_1) \dots T_u(\widehat{\boldsymbol{\vartheta}}_{S-1}, \overline{\boldsymbol{\vartheta}}_S) \\
& \times T_u(\overline{\boldsymbol{\vartheta}}_S, \check{\boldsymbol{\vartheta}}_{S-1}) \dots T_u(\check{\boldsymbol{\vartheta}}_1, \check{\boldsymbol{\vartheta}}_0)\, d\widehat{\boldsymbol{\vartheta}}_1 \dots d\overline{\boldsymbol{\vartheta}}_S \dots d\check{\boldsymbol{\vartheta}}_1.
\end{aligned}
$$

A.2 Posterior Distributions

A.2.1 Indicator Matrix

The indicator matrix $\boldsymbol{\Delta}$ can be sampled row-wise using Gibbs updates. The posterior probability that the m^{th} measurement is dedicated to the k^{th} factor (or not dedicated to any factor if $k = 0$) is a function of the marginal likelihood of its corresponding latent utility, for $k = 0, 1, \ldots, K$:

$$\Pr\big(\Delta_m = e_k \mid Y^\star_{\bullet m}, \boldsymbol{\Delta}_{-m}, \boldsymbol{X}, \boldsymbol{\theta}, \beta_m, \tau\big) = \frac{p(Y^\star_{\bullet m} \mid \Delta_m = e_k, \boldsymbol{X}, \boldsymbol{\theta}, \beta_m)\, p(\Delta_m = e_k \mid \boldsymbol{\Delta}_{-m}, \tau)}{\sum_{l=0}^{K} p(Y^\star_{\bullet m} \mid \Delta_m = e_l, \boldsymbol{X}, \boldsymbol{\theta}, \beta_m)\, p(\Delta_m = e_l \mid \boldsymbol{\Delta}_{-m}, \tau)}, \tag{A4}$$

where $p(Y^\star_{\bullet m} \mid \Delta_m = e_k, \boldsymbol{X}, \boldsymbol{\theta}, \beta_m)$ denotes the marginal likelihood of the vector $Y^\star_{\bullet m} = (Y^\star_{1m}, \ldots, Y^\star_{Nm})'$, conditioning on the remaining rows $\boldsymbol{\Delta}_{-m}$ of the indicator matrix.

From a computational point of view, these posterior probabilities can be calculated using the posterior log odds, which are more stable numerically than computing equation (A4) directly:

$$\Pr\big(\Delta_m = e_k \mid Y^\star_{\bullet m}, \boldsymbol{\Delta}_{-m}, \boldsymbol{X}, \boldsymbol{\theta}, \beta_m, \tau\big) = \left[\sum_{l=0}^{K} \exp\left(\mathcal{O}_{m,(k\to l)}\right)\right]^{-1},$$

where $\mathcal{O}_{m,(k\to l)}$ denotes the posterior log odds for a move from a model where measurement m is dedicated to factor k to a model where it is dedicated to factor l. More details on the posterior log odds are presented in Appendix A.3.

A.2.2 Idiosyncratic Variances and Factor Loadings

The idiosyncratic variances $\boldsymbol{\Sigma}$ corresponding to the continuous variables,[55] and the factor loadings $\boldsymbol{\alpha}$ are sampled simultaneously for each measurement m. Let $\alpha_m^{\boldsymbol{\Delta}}$ denote the only non-zero element in row m of the factor loading matrix, where the corresponding measurement m implicitly loads on factor k.

In the case of a continuous measurement that does not load on any factor ("null model" where $\Delta_m = e_0$), the idiosyncratic variance is sampled as follows:

$$\sigma_m^2 \mid \ldots \sim \mathcal{G}^{-1}\big(c_N;\, C_m^{Nn}\big),$$
$$c_N = c_0 + \frac{N}{2}, \qquad C_m^{Nn} = C_m^0 + \frac{\widetilde{Y}'_{\bullet m}\widetilde{Y}_{\bullet m}}{2},$$

[55] Recall that for the binary measurements, we set $\sigma_m^2 = 1$.

where $\widetilde{Y}_{\bullet m} = Y^{\star}_{\bullet m} - \boldsymbol{X}\beta_m$.

In the general case of a dedicated measurement, the posterior distributions of the idiosyncratic variance and of the non-zero factor loading are:

$$\sigma_m^2 \mid \ldots \sim \mathcal{G}^{-1}\big(c_N;\, C_m^N\big)\,, \qquad\qquad \alpha_m^{\boldsymbol{\Delta}} \mid \sigma_m^2, \ldots \sim \mathcal{N}\big(A_m^N a_m^N;\, A_m^N \sigma_m^2\big)\,, \tag{A5}$$

where, under the fixed-scale normal prior:

$$c_N = c_0 + \frac{N}{2}, \qquad\qquad C_m^N = C_m^0 + \frac{1}{2}\left(\widetilde{Y}'_{\bullet m}\widetilde{Y}_{\bullet m} + \frac{(a_m^0)^2}{A_m^0} - A_m^N (a_m^N)^2\right),$$
$$(A_m^N)^{-1} = (A_m^0)^{-1} + \sum_{i=1}^{N}\theta_{ik}^2, \qquad a_m^N = \frac{a_m^0}{A_m^0} + \sum_{i=1}^{N}\theta_{ik}\widetilde{Y}_{im}.$$

In the binary measurement case, non-zero factor loadings are sampled from equation (A5), where $\sigma_m^2 = 1$. No parameters need to be sampled in the "null model" case for binary measurements.

A.2.3 Regression Coefficients

The regression coefficients $\boldsymbol{\beta}$ are sampled row-wise from the following conditional distribution, for $m = 1, \ldots, M$:

$$\beta_m \sim \mathcal{N}\big(\boldsymbol{B}_m^N b_m^N;\, \boldsymbol{B}_m^N\big)\,,$$
$$(\boldsymbol{B}_m^N)^{-1} = (\boldsymbol{B}_m^0)^{-1} + \frac{1}{\sigma_m^2}\boldsymbol{X}'\boldsymbol{X}, \qquad b_m^N = (\boldsymbol{B}_m^0)^{-1} b_m^0 + \frac{1}{\sigma_m^2}\boldsymbol{X}'(Y^{\star}_{\bullet m} - \boldsymbol{\theta}\alpha_m),$$

where α_m is the column vector representing the m^{th} row of $\boldsymbol{\alpha}$.

A.2.4 Latent Variables for the Binary Measurements

If measurement m is dichotomous, its corresponding latent variable $Y^{\star}_{im}$ is sampled from the following truncated normal distribution, for each individual $i = 1, \ldots, N$:

$$Y^{\star}_{im} \sim \begin{cases} \mathcal{TN}_{(-\infty;0]}\left(X_i'\beta_m + \theta_i'\alpha_m;\, 1\right) & \text{if } Y_{im} = 0, \\ \mathcal{TN}_{(0;\infty)}\left(X_i'\beta_m + \theta_i'\alpha_m;\, 1\right) & \text{if } Y_{im} = 1. \end{cases}$$

A.2.5 Indicator Probabilities

The indicator probabilities τ are sampled by first drawing the components τ_0 and $\tau^\star$:

$$\begin{aligned} \tau_0 &\sim \mathcal{B}eta(\kappa_0 + n_0(\boldsymbol{\Delta});\, \xi_0 + M - n_0(\boldsymbol{\Delta}))\,, \qquad E(\tau_0) = \frac{\kappa_0 + n_0(\boldsymbol{\Delta})}{\kappa_0 + \xi_0 + M}, \\ \tau^\star &\sim \mathcal{D}ir(\kappa_1 + n_1(\boldsymbol{\Delta}), \ldots, \kappa_K + n_K(\boldsymbol{\Delta}))\,, \end{aligned} \tag{A6}$$

where $n_k(\boldsymbol{\Delta}) = \sum_{m=1}^{M} \mathbf{1}[\Delta_m = e_k]$ is the number of measurements dedicated to factor k (or not dedicated at all if $k = 0$). Then, compute the resulting probabilities τ using equation (14). In the case of the alternative hierarchical prior described in Subsection 2.3.1, each τ_{0m} is sampled from $\mathcal{B}eta(\kappa_0 + \mathbf{1}[\Delta_m = e_0]\,;\, \xi_0 + M - \mathbf{1}[\Delta_m = e_0])$. However, since only one observation is available for the update, it is recommended to integrate the τ parameters to obtain faster convergence and better mixing of the sampler (see Appendix A.3.2).

A.3 Posterior Log Odds

A.3.1 Deriving the Log Odds Conditional on the Indicator Probabilities τ

The posterior log odds that a measurement m currently dedicated to factor k becomes dedicated to factor k' ("null model" if k or $k' = 0$) can be expressed as:

$$\begin{aligned} \mathcal{O}_{m,(k\to k')} &= \log \frac{\Pr\Big(\Delta_m = e_{k'} \mid Y^\star_{\bullet m}, \boldsymbol{\Delta}_{-m}, \boldsymbol{X}, \boldsymbol{\theta}, \beta_m, \tau\Big)}{\Pr\Big(\Delta_m = e_k \mid Y^\star_{\bullet m}, \boldsymbol{\Delta}_{-m}, \boldsymbol{X}, \boldsymbol{\theta}, \beta_m, \tau\Big)}, \\ &= \log \frac{p(Y^\star_{\bullet m} \mid \Delta_m = e_{k'}, \boldsymbol{X}, \boldsymbol{\theta}, \beta_m)}{p(Y^\star_{\bullet m} \mid \Delta_m = e_k, \boldsymbol{X}, \boldsymbol{\theta}, \beta_m)} + \log \frac{p(\Delta_m = e_{k'}, \boldsymbol{\Delta}_{-m} \mid \tau)}{p(\Delta_m = e_k, \boldsymbol{\Delta}_{-m} \mid \tau)}, \\ &= -\mathcal{O}_{m,(k'\to k)}, \end{aligned} \tag{A7}$$

where the last term is equal to $\log(\tau_{k'}/\tau_k)$ when sampling is done conditional on the parameters τ (see Appendix A.3.2 for the case where τ is integrated out).

The marginal likelihoods of the latent variables of the measurements are required to compute the posterior log odds. These marginal likelihoods differ for continuous and binary measurements, and for the case of the "null model" and the general case of a dedicated measurement. In the continuous case, they can be expressed as, using the posterior moments

derived in A.2.2:

$$p(Y^{\star}_{\bullet m} \mid \Delta_m, \boldsymbol{X}, \boldsymbol{\theta}, \beta_m) = \begin{cases} (2\pi)^{-\frac{N}{2}} \dfrac{\Gamma(c_N)\,(C_m^0)^{c_0}}{\Gamma(c_0)\,(C_m^{Nn})^{c_N}} & \text{in the "null model,"} \\ (2\pi)^{-\frac{N}{2}} \dfrac{|A_m^N|^{1/2}}{|A_m^0|^{1/2}} \dfrac{\Gamma(c_N)\,(C_m^0)^{c_0}}{\Gamma(c_0)\,(C_m^N)^{c_m^N}} & \text{in the dedicated case.} \end{cases}$$

while in the binary case:

$$p(Y^{\star}_{\bullet m} \mid \Delta_m, \boldsymbol{X}, \boldsymbol{\theta}, \beta_m) = \begin{cases} (2\pi)^{-\frac{N}{2}} \exp\left\{-\frac{1}{2}\widetilde{Y}'_{\bullet m}\widetilde{Y}_{\bullet m}\right\} & \text{in the "null model," otherwise} \\ (2\pi)^{-\frac{N}{2}} \dfrac{|A_m^N|^{1/2}}{|A_m^0|^{1/2}} \exp\left\{-\frac{1}{2}\left(\widetilde{Y}'_{\bullet m}\widetilde{Y}_{\bullet m} + \dfrac{(a_m^0)^2}{A_m^0} - A_m^N (a_m^N)^2\right)\right\}. \end{cases}$$

With these marginal likelihoods in hand, it is straightforward to compute the posterior log odds. In the continuous measurement case, they are equal to:[56]

$$\begin{aligned} \mathcal{O}_{m,(0\to k)} &= -\frac{1}{2}\log(P_{mk}) - c_N \log\left(1 - \frac{Q_{mk}}{C_m^{Nn}}\right) + \log\frac{\tau_k}{\tau_0}, \\ \mathcal{O}_{m,(k\to k')} &= -\frac{1}{2}\log\left(\frac{P_{mk'}}{P_{mk}}\right) - c_N \log\left(\frac{C_m^{Nn} - Q_{mk'}}{C_m^{Nn} - Q_{mk}}\right) + \log\frac{\tau_{k'}}{\tau_k}, \end{aligned}$$

and in the binary case:

$$\begin{aligned} \mathcal{O}_{m,(0\to k)} &= -\frac{1}{2}\log(P_{mk}) + Q_{mk} + \log\frac{\tau_k}{\tau_0}, \\ \mathcal{O}_{m,(k\to k')} &= -\frac{1}{2}\log\left(\frac{P_{mk'}}{P_{mk}}\right) + Q_{mk'} - Q_{mk} + \log\frac{\tau_{k'}}{\tau_k}, \end{aligned}$$

for all $k \neq 0$ and $k' \neq 0$, where:

$$P_{mk} = 1 + A_m^0 \sum_{i=1}^{N} \theta_{ik}^2, \qquad\qquad Q_{mk} = \frac{1}{2}\frac{\left(\sum_{i=1}^{N} \theta_{ik}\widetilde{Y}_{im}\right)^2}{\left((A_m^0)^{-1} + \sum_{i=1}^{N}\theta_{ik}^2\right)}.$$

A.3.2 Integrating out the Indicator Probabilities τ

Integrating out the vector of indicator probabilities τ from the likelihood function does not affect the ratio of the marginal likelihoods of $Y^{\star}_{\bullet m}$ in equation (A7), but only the last term that is equal to $\log(\tau_{k'}/\tau_k)$ when sampling is done conditional on τ. For a move from a model

[56] For the computation of the posterior log odds, the factor loadings are assumed to be *a priori* centered (i.e., $a_m^0 = 0$) to simplify the calculations. This assumption is usually adopted in factor analysis.

where measurement m is dedicated to factor k to a model where it is dedicated to factor k', this second term should be replaced by the ratio of the marginal likelihoods of $\boldsymbol{\Delta}$ in the two models. This ratio is expressed as $\log(p(\boldsymbol{\Delta}^{mk'})/p(\boldsymbol{\Delta}^{mk}))$, where the two indicator matrices $\boldsymbol{\Delta}^{mk}$ and $\boldsymbol{\Delta}^{mk'}$ are identical up to row m, where in the first case this row is the indicator vector e_k, while in the second it is $e_{k'}$.

The marginal distribution of the indicator matrix $\boldsymbol{\Delta}$ is equal to:

$$\begin{aligned} p(\boldsymbol{\Delta}) &= \iint p(\boldsymbol{\Delta} \mid \tau_0, \tau^\star) p(\tau_0) p(\tau^\star) d\tau_0 d\tau^\star, \\ &= \frac{B(\kappa_0 + n_0(\boldsymbol{\Delta}), \xi_0 + M - n_0(\boldsymbol{\Delta}))}{B(\kappa_0, \xi_0)} \frac{\Gamma\Big(\sum_{k=1}^K \kappa_k\Big) \prod_{k=1}^K \Gamma(\kappa_k + n_k(\boldsymbol{\Delta}))}{\Gamma\Big(M - n_0(\boldsymbol{\Delta}) + \sum_{k=1}^K \kappa_k\Big) \prod_{k=1}^K \Gamma(\kappa_k)}, \end{aligned} \tag{A8}$$

where $n_k(\boldsymbol{\Delta}) = \sum_{m=1}^M \mathbf{1}[\Delta_m = e_k]$, for $k = 0, \ldots, K$. Given that the numbers of measurements in the different groups are equal to:

$$\begin{aligned} n_k(\boldsymbol{\Delta}^{mk'}) &= n_k(\boldsymbol{\Delta}^{mk}) - 1, & n_l(\boldsymbol{\Delta}^{mk'}) &= n_l(\boldsymbol{\Delta}^{mk}), \\ n_{k'}(\boldsymbol{\Delta}^{mk'}) &= n_{k'}(\boldsymbol{\Delta}^{mk}) + 1, \end{aligned}$$

for all $k \neq k'$ and $l \notin \{k, k'\}$, it follows from equation (A8) that the ratio of the marginal likelihoods, for all $k \neq k'$, simplifies to:

$$\frac{p(\boldsymbol{\Delta}^{mk'})}{p(\boldsymbol{\Delta}^{mk})} = \begin{cases} \dfrac{n_{k'}(\boldsymbol{\Delta}^{mk}) + \kappa_{k'}}{n_k(\boldsymbol{\Delta}^{mk}) - 1 + \kappa_k} & \text{for } k \neq 0 \text{ and } k' \neq 0, \\ \dfrac{n_0(\boldsymbol{\Delta}^{mk}) + \kappa_0}{n_k(\boldsymbol{\Delta}^{mk}) - 1 + \kappa_k} \dfrac{M - n_0(\boldsymbol{\Delta}^{mk}) - 1 + \sum_{l=1}^K \kappa_l}{M - n_0(\boldsymbol{\Delta}^{mk}) - 1 + \xi_0} & \text{for } k' = 0. \end{cases}$$

In the case of the alternative hierarchical prior specification on τ, with individual τ_{0m} parameters but common $\tau^\star$ for the measurements, the marginal distribution of the indicator matrix $\boldsymbol{\Delta}$ is:

$$p(\boldsymbol{\Delta}) = \frac{(\kappa_0)^{n_0(\boldsymbol{\Delta})} (\xi_0)^{M - n_0(\boldsymbol{\Delta})}}{(\kappa_0 + \xi_0)^M} \frac{\Gamma\Big(\sum_{k=1}^K \kappa_k\Big) \prod_{k=1}^K \Gamma(\kappa_k + n_k(\boldsymbol{\Delta}))}{\Gamma\Big(M - n_0(\boldsymbol{\Delta}) + \sum_{k=1}^K \kappa_k\Big) \prod_{k=1}^K \Gamma(\kappa_k)},$$

and the ratio of marginal likelihoods, for a move from k to k' in row m, for $k \neq k'$, simplifies to:

$$\frac{p(\boldsymbol{\Delta}^{mk'})}{p(\boldsymbol{\Delta}^{mk})} = \begin{cases} \dfrac{n_{k'}(\boldsymbol{\Delta}^{mk}) + \kappa_{k'}}{n_k(\boldsymbol{\Delta}^{mk}) - 1 + \kappa_k} & \text{for } k \neq 0 \text{ and } k' \neq 0, \\ \dfrac{\kappa_0}{\xi_0} \dfrac{M - n_0(\boldsymbol{\Delta}^{mk}) - 1 + \sum_{l=1}^{K} \kappa_l}{n_k(\boldsymbol{\Delta}^{mk}) - 1 + \kappa_k} & \text{for } k' = 0. \end{cases} \tag{A9}$$

B Data: The British Cohort Study

We use data from the British Cohort Study (BCS), a survey of all babies born (alive or dead) after the 24$^{\text{th}}$ week of gestation from 00.01 hours on Sunday, 5$^{\text{th}}$ April to 24.00 hours on Saturday, April 11$^{\text{th}}$, 1970 in England, Scotland, Wales and Northern Ireland. There have been seven follow-ups on the members of the birth cohort: in 1975, 1980, 1986, 1996, 2000, 2004 and 2008. We draw information on background characteristics from the birth survey, and on cognitive, mental and physical health measurements from the second sweep (age 10). We exclude children born with congenital abnormalities, non-whites, and respondents with missing information on the background characteristics. Individuals with missing observations on some of the cognitive, mental and physical health measurements are discarded from the sample, so we are left with a sample of 2,080 men.

B.1 The Measurement System

The measurement system includes one hundred and thirty one indicators of child cognitive, mental and physical health traits, all collected at age ten. Notice we use both binary and continuous measurements, which have been standardized to have zero mean and standard deviation equal to one for use in our empirical application.

Cognitive Ability Scales. As indicators of cognitive ability, we use the following seven tests:

- The Picture Language Comprehension Test [PLCT]: this is a new test specifically developed for the BCS on the basis of the American Peabody Picture Vocabulary Test and the English Picture Vocabulary Test; it covers vocabulary, sequence and sentence comprehension.
- The Friendly Math Test [FMT]: this is a new test specifically designed for the BCS; it covers arithmetic, fractions, algebra, geometry and statistics.

- The Shortened Edinburgh Reading Test [SERT]: this is a shortened version of the Edinburgh Reading Test, which is a test of word recognition particularly designed to capture poor readers; it covers vocabulary, syntax, sequencing, comprehension, and retention.

- The four British Ability Scales [BAS]: these measure a construct similar to IQ, and include two non-verbal scales (Matrices [**BASTM**] and Recall Digits [**BASTRD**]) and two verbal scales (Similarities [WS] and Word Definition [**BASTWD**]).

Mental Health Scales. As indicators of psychological and behavioral problems, we use the items from the following five tests:

1. The Rutter Parental 'A' Scale of Behavioral Disorder (Rutter et al., 1970): it was administered to the mother, and designed to capture the presence of problem behaviors. It contains 19 items which are descriptions of behavior, and the mother was asked to indicate whether each description 'does not apply', 'applies somewhat' or 'definitely applies' to the child, on a scale from 0 to 100. A visual analogue scale was used: the mother had to draw a vertical line through the printed horizontal line to show how much a behavior applied (or not) to the child.

2. The Conners Hyperactivity Scale (Conners, 1969): it was also administered to the mother, and developed to assess attention deficit/hyperactivity disorder and evaluate problem behavior in children and adolescents. It includes 19 items, and the mother was asked to indicate whether each description applied to the child on a scale from 0 to 100, using a visual analogue like for the Rutter Scale.

3. The Child Developmental Scale: it was administered to a teacher with knowledge of the child, to assess the child neurodevelopmental behavior against the 'average' behavior of most children of a similar age. It includes 53 items, and the teachers were asked to indicate their level of agreement with each statement by bisecting a line, which was coded into a 47-point scale ranging from "Not at all" to "A great deal". The items for this scale were taken mainly from the Conners Teachers Hyperactivity Rating Scale (Conners, 1969) and the Rutter Teacher Behavioral Scale 'B' (Rutter, 1967), and questions from the Swansea Assessment Battery (with permission of Professor Maurice Chazan; see Butler et al., 1997).

4. The Self-Esteem (Lawseq) Scale: it was administered by the teacher and completed by the child to measure his self-esteem with reference to teachers, peers and parents. It includes 12 items (The total number of questions included is 16, but four of them

are distractors) and was created by former Chief Educational Psychologist of Somerset LEA (Local Education Authority) Lawrence (Lawrence, 1973, 1978). The child was asked to answer 'yes', 'no' or 'don't know', where the answer 'no' represents a higher level of self-esteem (only one question is reverse-scored, and we have recoded it accordingly). For use in our empirical application, we have recoded all the answers into binary measurements, by giving a value of 1 to all the 'no' answers, and a value of 0 to all the 'yes' and 'don't know' answers.

5. The Locus of Control (Caraloc) Scale: it was administered by the teacher and completed by the child to measure his perceived achievement control. It includes 16 items (the total number of questions included is 20, but four of them are distractors) and was constructed from several well known tests of locus of control (Gammage, 1975). The child was asked to answer 'yes', 'no' or 'don't know', where the answer 'no' represents a more internal locus of control (only one question is reverse-scored, and we have recoded it accordingly), which is desirable and also referred to as "self-agency", "personal control", "self-determination", etc. For use in our empirical application, we have recoded all the answers into binary measurements, by giving a value of 1 to all the 'no' answers, and a value of 0 to all the 'yes' and 'don't know' answers (a similar scoring scheme has been used in Ternouth et al., 2009).

Physical Health. As indicators of physical health, we use the following five measures, all recorded during medical examinations: height, head circumference, weight, systolic and diastolic blood pressure.[57]

B.2 Control Variables

The following seven control variables — denoted $\boldsymbol{X}$ in our model — are included in the measurement system. The variables have been standardized to have zero mean and standard deviation equal to one. i) mother's age at birth, ii) mother's education at birth (a dummy variable for whether the mother continued education beyond the minimum school-leaving

[57] While the availability of information on height and weight is not a unique feature of our data, differently from our case most of the measures recorded in public-use data are self-reported: as such, they are subject to substantial measurement error, which is usually not dealt with by researchers with the use of suitable methods such as factor-analytic techniques as we instead do here.

age[58]), iii) father's high social class at birth,[59] iv) total gross family income at age 10,[60] v) an indicator for broken family (a dummy variable for whether the child lived with both parents since birth until age 10), vi) the number of previous livebirths, and vii) the number of children in the family at age 10.

[58]The compulsory minimum school leaving age was increased from fourteen to fifteen in 1947, following the 1944 Education Act.

[59]The BCS uses the Registrar General's classification for measuring social class (SC). High Social Class comprises SCI, SCII and SCIIINM (Non-Manual). Social class I includes professionals, such as lawyers, architects and doctors; Social Class II includes intermediate workers, such as shopkeepers, farmers and teachers; Social Class III Non Manual includes skilled non-manual workers, such as shop assistants and clerical workers in offices.

[60]This is a categorical indicator taking the following values: 1=under £35 pw; 2=£35-49 pw; 3=£50-99 pw; 4=£100-149 pw; 5=£150-199 pw; 6=£200-249 pw; 7=£250 or more per week.

References

Achenbach, T. M., S. H. McConaughy, and C. T. Howell (1987): "Child/adolescent behavioral and emotional problems: implications of cross-informant correlations for situational specificity." *Psychological bulletin*, 101, 213.

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

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

Antecol, H. and D. Cobb-Clark (2010): "Do Non-Cognitive Skills Help Explain the Occupational Segregation of Young People?" *IZA Discussion Paper No. 5093*.

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

Barnett, W. A. and M. Chauvet (2011): *Financial aggregation and index number theory*, vol. 2 of *Surveys on theories in economics and business administration*, Hackensack, NJ: World Scientific Publishing.

Baron, J. and D. Cobb-Clark (2010): "Are young people's educational outcomes linked to their sense of control?" *IZA Discussion Paper No. 4907*.

Bekker, P. A. and J. M. ten Berge (1997): "Generic Global Indentification in Factor Analysis," *Linear Algebra and its Applications*, 264, 255–263.

Berger, J. O. (1985): *Statistical Decision Theory and Bayesian Analysis*, New York: Springer-Verlag, 2 ed.

Bhattacharya, A. and D. Dunson (2011): "Sparse Bayesian Infinite Factor Models," *Biometrika*, 98, 291–306.

Blanden, J., P. Gregg, and L. Macmillan (2007): "Accounting for intergenerational income persistence: noncognitive skills, ability and education," *Economic Journal*, 117, C43–C60.

Bonhomme, S. and J. Robin (2010): "Generalized non-parametric deconvolution with an application to earnings dynamics," *Review of Economic Studies*, 77, 491–533.

Browne, M. W. (2001): "An overview of analytic rotation in exploratory factor analysis," *Multivariate Behavioral Research*, 36, 111–150.

Butler, N., S. Despotidou, and P. Shepherd (1997): "British Cohort Study (BCS70) ten-year follow-up: A guide to the BCS70 10-year data available at the Economic and Social Research Unit data archive," Working paper, London: Social Statistics Research Unit, City University.

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, 361–422.

Carroll, J. (1953): "An analytical solution for approximating simple structure in factor analysis," *Psychometrika*, 18, 23–38, 10.1007/BF02289025.

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, 1438–1456.

Casella, G. and C. P. Robert (2004): *Monte Carlo Statistical Methods*, Springer, 2nd ed.

Cattell, R. (1966): "The data box: its ordering of total resources in terms of possible relational systems. 67-128 in: RB Cattell," *Handbook of multivariate experimental psychology*.

Cattell, R. B. (1952): *Factor analysis: An introduction and manual for the psychologist and social scientist*, Harper.

Chen, M., A. Zaas, C. Woods, G. S. Ginsburg, J. Lucas, D. Dunson, and L. Carin (2011): "Predicting Viral Infection From High-Dimensional Biomarker Trajectories," *Journal of the American Statistical Association*, 106, 1259–1279.

Chib, S. and E. Greenberg (1995): "Understanding the Metropolis-Hastings Algorithm," *The American Statistician*, 49, 327–335.

Conners, C. K. (1969): "A Teacher Rating Scale for Use in Drug Studies with Children," *The American Journal of Psychiatry*, 126, 884–888.

Conti, G., J. J. Heckman, and S. Urzúa (2010): "The Education-Health Gradient," *American Economic Review: Papers and Proceedings*, 100, 1–5.

Costa, P. T. and R. R. McCrae (1992): "Revised NEO personality inventory (NEO-PI-R) and NEO five-factor (NEO-FFI) inventory professional manual," *Odessa, FL: PAR*.

Costa, P. T., R. R. McCrae, and D. A. Dye (1991): "Facet scales for agreeableness and conscientiousness: a revision of tshe NEO personality inventory," *Personality and Individual Differences*, 12, 887–898.

Costello, A. B. and J. W. Osborne (2005): "Best Practices in Exploratory Factor Analysis: Four Recommendations for Getting the Most From Your Analysis," *Practical Assessment, Research & Evaluation*, 10.

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

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

Dohmen, T., A. Falk, D. Huffman, and U. Sunde (2012): "The intergenerational transmission of risk and trust attitudes," *The Review of Economic Studies*, 79, 645–677.

Fabrigar, L. R., D. T. Wegener, R. C. MacCallum, and E. J. Strahan (1999): "Evaluating the use of exploratory factor analysis in psychological research," *Psychological methods*, 4, 272–299.

Feinstein, L. (2000): "The Relative Economic Importance of Academic, Psychological and Behavioural Attributes Developed on Chilhood," *CEP Discussion Paper 443*.

Ferguson, G. A. (1954): "The concept of parsimony in factor analysis," *Psychometrika*, 19, 281–290.

Fiorini, M. and P. M. Keane (2012): "How the allocation of children's time affects cognitive and non-cognitive development," *Working Paper*.

Frühwirth-Schnatter, S. (2006): *Finite Mixture and Markov Switching Models*, Springer.

Frühwirth-Schnatter, S. and H. F. Lopes (2012): "Parsimonious Bayesian Factor Analysis When the Number of Factors Is Unknown," Unpublished Technical Report.

Gale, C. R., G. D. Batty, and I. J. Deary (2008): "Locus of control at age 10 years and health outcomes and behaviors at age 30 years: the 1970 British Cohort Study," *Psychosomatic Medicine*, 70, 397–403.

Gale, C. R., S. L. Hatch, G. D. Batty, and I. J. Deary (2009): "Intelligence in childhood and risk of psychological distress in adulthood: The 1958 National Child Development Survey and the 1970 British Cohort Study," *Intelligence*, 37, 592–599.

Gammage, P. (1975): *Socialization, schooling and locus of control*, Bristol University, PhD Thesis.

George, E. I. and R. E. McCulloch (1997): "Approaches for Bayesian Variable Selection," *Statistica Sinica*, 7, 339–373.

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

Geweke, J. F. (1996): "Variable Selection and Model Comparison in Regresssion," in *Bayesian Statistics 5*, ed. by J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, Oxford: Oxford University Press, 609–620.

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

Ghahramani, Z., T. L. Griffiths, and P. Sollich (2007): "Bayesian Nonparametric Latent Feature Models," in *Bayesian Statistics*, Oxford University Press, 201–225.

Goldberg, L. R. (1990): "An Alternative "Description of Personality": The Big-Five Factor Structure," *Journal of Personality and Social Psychology*, 59, 1216–1229.

Gorsuch, R. (1983): *Factor Analysis*, Hillsdale, NJ: Lawrence Erlbaum Associates.

Gorsuch, R. L. (2003): "Factor Analysis," in *Handbook of psychology: Research methods in psychology*, ed. by I. B. Weiner, D. K. Freedheim, J. A. Schinka, and W. F. Velicer, Hoboken, NJ: John Wiely & Sons, Inc., vol. 2, chap. 6, 143–164.

Green, P. J. (1995): "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination," *Biometrika*, 82, 711–732.

Hastings, W. K. (1970): "Monte Carlo Sampling Methods Using Markov Chains and Their Applications," *Biometrika*, 57, 97–109.

Heckman, J. J., R. Pinto, and P. A. Savelyev (2013): "Understanding the Mechanisms Through Which an Influential Early Childhood Program Boosted Adult Outcomes," *American Economic Review*, 103, 1–35.

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, 411–482.

Helmers, C. and M. Patnam (2011): "The formation and evolution of childhood skill acquisition: Evidence from India," *Journal of Development Economics*, 95, 252–266.

Heywood, H. B. (1931): "On Finite Sequences of Real Numbers," *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 134, 486–501.

Hofmann, R. J. (1978): "Complexity and simplicity as objective indices descriptive of factor solutions," *Multivariate Behavioral Research*, 13, 247–250.

Huang, A. and M. P. Wand (2013): "Simple marginally noninformative prior distributions for covariance matrices," *Bayesian Analysis*, 8, 439–452.

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

Jennrich, R. I. (2001): "A Simple General Procedure for Orthogonal Rotation," *Psychometrika*, 66, 289–306.

——— (2002): "A Simple General Method for Oblique Rotation," *Psychometrika*, 67, 7–20.

——— (2004): "Rotation to Simple Loadings using Component Loss Functions: The Orthogonal Case," *Psychometrika*, 69, 257–273.

——— (2006): "Rotation to Simple Loadings using Component Loss Functions: The Oblique Case," *Psychometrika*, 71, 173–191.

——— (2007): "Rotation Algorithms: From Beginning to End," in *Handbook of Latent Variable and Related Models*, ed. by Sik-Yum Lee, North-Holland/Elsevier, chap. 3, 45–63.

Jolliffe, I. T. (1986): *Principal component analysis*, vol. 487, Springer-Verlag New York.

Jones, A. M., N. Rice, and P. R. Dias (2011): "Long-term effects of school quality on health and lifestyle: Evidence from comprehensive schooling reforms in England," *Journal of Human Capital*, 5, 342–376.

Kaestner, R. (2009): "Adolescent Cognitive and Non-cognitive Correlates of Adult Health," Tech. rep., National Bureau of Economic Research.

Kastner, G. and S. Frühwirth-Schnatter (2014): "Ancillarity-Sufficiency Interweaving Strategy (ASIS) for Boosting MCMC Estimation of Stochastic Volatility," *Computational Statistics & Data Analysis*, 76, 408–423.

Knowles, D. and Z. Ghahramani (2007): "Infinite Sparse Factor Analysis and Infinite Independent Components Analysis," in *Independent Component Analysis and Signal Separation*, ed. by M. E. Davies, C. J. James, S. A. Abdallah, and M. D. Plumbley, Berlin Heidelberg: Springer, vol. 4666 of *Lecture Notes in Computer Science*, 381–388.

Lawrence, D. (1973): "Improved reading through counselling," Tech. rep., London: Ward Lock.

——— (1978): "Counselling students with reading difficulties: A handbook for tutors and organisors," Tech. rep., London: Good Reading.

Le Cam, L. (1986): *Asymptotic Methods in Statistical Decision Theory*, New York: Springer.

Ledermann, W. (1937): "On the Rank of the Reduced Correlational Matrix in Multiple-Factor Analysis," *Psychometrika*, 2, 85–93.

Liu, J. S., W. H. Wong, and A. Kong (1994): "Covariance Structure of the Gibbs Sampler with Applications to the Comparisons of Estimators and Augmentation Schemes," *Biometrika*, 81, 27–40.

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

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

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, 897–914.

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

Lucas, J., 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, Cambridge University Press, 155–176.

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

Mervielde, I., V. Buyst, and F. De Fruyt (1995): "The validity of the Big-five as a model for teachers' ratings of individual differences among children aged 4–12 years," *Personality and Individual Differences*, 18, 525–534.

Murasko, J. E. (2007): "A lifecourse study on education and health: The relationship between childhood psychosocial resources and outcomes in adolescence and young adulthood," *Social Science Research*, 36, 1348–1370.

Natarajan, R. and C. E. McCulloch (1998): "Gibbs Sampling With Diffuse Proper Priors: A Valid Approach to Data-Driven Inference?" *Journal of Computational and Graphical Statistics*, 7, 267–277.

Neal, R. M. (1996): "Sampling from Multimodal Distributions Using Tempered Transitions," *Statistics and Computing*, 6, 353–366.

Onatski, A. (2009): "Testing hypotheses about the number of factors in large factor models," *Econometrica*, 77, 1447–1479.

Paisley, J. and L. Carin (2009): "Nonparametric Factor Analysis with Beta Process Priors," *Proceedings of the 26th Annual International Conference on Machine Learning*, 1–8.

Poirier, D. J. (1998): "Revising Beliefs in Nonidentified Models," *Econometric Theory*, 14, 483–509.

R Core Team (2013): *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

Richardson, S. and P. J. Green (1997): "On Bayesian analysis of mixtures with an unknown number of components," *Journal of the Royal Statistical Society. Series B (Methodological)*, 59, 731–792.

Rutter, M. (1967): "A Children's Behaviour Questionnaire for Completion by Teachers: Preliminary Findings," *Journal of Child Psychology and Psychiatry*, 8, 1–11.

Rutter, M., J. Tizard, and K. Whitmore (1970): *Education, health and behaviour*, London, UK: Longmans.

Sato, M. (1992): "A Study of an Identification Problem and Substitute Use of Principal Component Analysis in Factor Analysis," *Hiroshima Mathematical Journal*, 22, 479–524.

Saunders, D. R. (1953): "An analytic method for rotation to orthogonal simple structure," *Research Bulletin*, 53.

Stephens, M. (2000): "Bayesian Analysis of Mixture Models with an Unknown Number of Components - An Alternative to Reversible Jump Methods," *The Annals of Statistics*, 28, 40–74.

Tabachnick, B. G. and L. S. Fidell (2001): *Using multivariate statistics*, Boston: Allyn and Bacon, chap. Principal components and factor analysis, 582–633, 4th ed.

Ternouth, A., D. Collier, and B. Maughan (2009): "Childhood emotional problems and self-perceptions predict weight gain in a longitudinal regression model," *BMC medicine*, 7, 7–46.

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

——— (1947): *Multiple factor analysis*, Chicago: University of Chicago Press.

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

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

Van Lieshout, C. and G. Haselager (1994): "The Big Five personality factors in Q-sort descriptions of children and adolescents," *The developing structure of temperament and personality from infancy to adulthood*, 293–318.

von Stumm, S., C. R. Gale, G. D. Batty, and I. J. Deary (2009): "Childhood intelligence, locus of control and behaviour disturbance as determinants of intergenerational social mobility: British Cohort Study 1970," *Intelligence*, 37, 329–340.

West, M. (2003): "Bayesian Factor Regression Models in the 'Large p, Small n' Paradigm," in *Bayesian Statistics 7*, ed. by J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West, Oxford: Oxford University Press, vol. 7, 733–742.

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, 880–896.