Journal of Econometrics ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Bayesian exploratory factor analysis☆

Gabriella Conti[a], Sylvia Frühwirth-Schnatter[b], James J. Heckman[c,d], Rémi Piatek[e,*]

[a] Department of Applied Health Research, University College London, UK
[b] Department of Finance, Accounting, and Statistics, Vienna University of Economics and Business, Austria
[c] Department of Economics, University of Chicago, USA
[d] American Bar Foundation, USA
[e] Department of Economics, University of Copenhagen, Denmark

ARTICLE INFO

Article history:
Available online xxxx

JEL classification:
C11
C38
C63

Keywords:
Bayesian factor models
Exploratory factor analysis
Identifiability
Marginal data augmentation
Model expansion
Model selection

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.

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 (EFA) is a well developed classical procedure for doing dedicated factor analysis (Gorsuch, 1983, 2003). The various steps required in executing classical 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

☆ A Web Appendix containing additional material is available at http://heckman.uchicago.edu/BayesFA.

* Correspondence to: Department of Economics, University of Copenhagen, Øster Farimagsgade 5, DK-1353 Copenhagen K, Denmark. Tel.: +45 35 32 30 35.
E-mail address: Remi.Piatek@econ.ku.dk (R. Piatek).

http://dx.doi.org/10.1016/j.jeconom.2014.06.008
0304-4076/

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). Unlike 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 modeling, 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 $Y^\star_{im}$:

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

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

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

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 $\boldsymbol{Y}^\star = (Y^\star_1, \ldots, Y^\star_N)'$ 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 measurements 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

[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.

Please cite this article in press as: Conti, G., et al., Bayesian exploratory factor analysis. Journal of Econometrics (2014), http://dx.doi.org/10.1016/j.jeconom.2014.06.008