Computational Statistics and Data Analysis 99 (2016) 189–203

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

A new method for simultaneous estimation of the factor model parameters, factor scores, and unique parts

Alwin Stegeman *

Heymans Institute for Psychological Research, University of Groningen, Grote Kruisstraat 2/1, 9712 TS Groningen, The Netherlands

ARTICLE INFO

Article history:
Received 14 July 2015
Received in revised form 10 January 2016
Accepted 20 January 2016
Available online 10 February 2016

Keywords:
Common factor analysis
Simultaneous estimation
Direct fitting
Minimum rank factor analysis
Factor indeterminacy

ABSTRACT

In the common factor model the observed data is conceptually split into a common covariance producing part and an uncorrelated unique part. The common factor model is fitted to the data itself and a new method is introduced for the simultaneous estimation of loadings, unique variances, factor scores, and unique parts. The method is based on Minimum Rank Factor Analysis and allows for the percentage of explained common variance to be computed. Taking into account factor indeterminacy, an explicit description of the complete class of solutions for the factor scores and unique parts is given. The method is evaluated in a simulation study and fitted to a dataset in the literature.

1. Introduction

We consider exploratory factor analysis of continuous observed variables. In common factor analysis the observed variables are conceptually split up into a common part and a unique part. The common part of each variable is correlated to other observed variables, while the unique part is not. The latter includes measurement error and possibly a specific part uniquely measured by the corresponding observed variable. The common part of all observed variables is approximated by a small number of underlying latent factors. The origins of the common factor model (or factor analysis model) date back to Spearman (1904) and Thurstone (1935). A distinction is made in the literature between the *random factor model*, in which the factors and unique parts are considered random variables, and the *fixed factor model*, in which only the unique parts are random variables and the factor scores are parameters to be estimated. Fitting of the factor model is commonly done on the observed covariance or correlation matrix, with the loadings, unique variances, and factor correlations as parameters to be estimated. For the fixed factor model the factor scores are commonly estimated in a second step, using a (weighted) least squares criterion.

In this paper, we introduce a factor model that addresses two problematic issues in the foundation and application of common factor analysis. The first issue concerns the measures that are used to assess the fit of the factor model. In practice, the fit is assessed by comparing the observed correlation matrix to that of the estimated factor model, including the unique variances. This is done by, e.g., least squares (Harman and Jones, 1966), or a chi-square measure in a maximum likelihood framework (Jöreskog, 1967). However, no distinction is made between the common variances to be explained and the common variances in the estimated factor model. The latter are produced as "communalities" by statistical software packages, thus suggesting that 100% of the common variance is explained. But in practice we have imperfect fit of the factor

* Tel.: +31 50 363 6193; fax: +31 50 363 6304.
E-mail address: a.w.stegeman@rug.nl.
URL: http://www.alwinstegeman.nl.

http://dx.doi.org/10.1016/j.csda.2016.01.012

model and not all common variance will be explained. The question remains how well the latent factors explain the common variance in the observed variables, which is important when choosing the number of factors. The only estimation method that provides an answer to this question and yields a percentage of explained common variance is Minimum Rank Factor Analysis (MRFA) of Ten Berge and Kiers (1991). See also Ten Berge (1998).

The second issue concerns the nonuniqueness of factor scores under the factor model, known in the literature as *factor indeterminacy* and first described by Wilson (1928). Factor indeterminacy is a fundamental property of the common factor model and occurs because the model contains more factors, including unique parts, than observed variables. Guttman (1955) showed that under perfect fit the factor scores (and unique parts) can be written as the sum of a determinate part and an orthogonal indeterminate part. The determinate part is the regression of the vector of factor scores (or the unique part) on the observed variables, while the indeterminate part represents all possible residuals of this regression. Moreover, Guttman (1955) proved a tight lower bound on the correlation between alternative factors (with the same determinate but different indeterminate parts). When this minimal correlation is close to zero, the indeterminacy of the factor scores also affects the interpretation of the factor. Indeed, how could nearly uncorrelated factors represent the same latent trait? Although factor indeterminacy may have serious consequences for factor interpretation and factor score estimation, it is usually ignored in applications and minimal correlations are not reported (Steiger, 1979; Maraun, 1996). In the empirical Bayes framework for factor score estimation (which implies using unweighted least squares) the uncertainty in the factor scores due to indeterminacy can be computed explicitly as the posterior variance of the factor scores given the observed data and estimates of the loadings, unique variances, and factor correlations (Bartholomew, 1981; Skrondal and Rabe-Hesketh, 2004). However, this implies that factors are treated as random variables which is seen by some as contradicting the idea that factor scores are parameters to be estimated (Bartholomew, 1981). In the literature, factor indeterminacy remains a highly controversial topic and experts do not seem to agree on how to address the issue (Maraun, 1996). For an overview see also Mulaik (2010).

In this paper, we present a novel model and estimation procedure of common factor analysis in which the loadings, unique variances, factor scores, and unique parts are all parameters and are estimated simultaneously. We refer to this type of factor model as the *data factor model*. Our model is a constrained version of the data factor model introduced independently by Kiers in Sočan (2003) and by De Leeuw (2004), and combines MRFA with unweighted least squares estimation of factor scores and unique parts. As a result, the explained common variance can be computed. Moreover, we extend the analysis of factor indeterminacy of Guttman (1955) to the case of imperfect fit and apply it to our model and that of De Leeuw (2004). We obtain a description of the complete class of solutions for the factor scores and unique parts, decomposed into determinate and indeterminate parts, and an expression for the minimal correlation between alternative factors. The unweighted least squares factor score estimate, its sampling variance, and its variance due to indeterminacy equal their empirical Bayes counterparts. The description of the complete class of solutions for the factor scores makes it possible to obtain probability densities of factor scores due to indeterminacy by sampling the indeterminate parts. The sampling is done uniformly and without assuming a distributional form for the factor scores. These probability densities can be used in practice as a visual tool to aid the researcher in deciding how to value the obtained factor score estimates in the presence of indeterminacy. In the empirical Bayes framework such probability densities can also be obtained, but only after specifying a prior distribution for the factor scores (Bartholomew, 1981).

Large sample theory under normality has yet been derived for the data factor model of Sočan (2003) and De Leeuw (2004), but the asymptotic distribution of the unique variance estimates by MRFA can be found in Shapiro and Ten Berge (2002). However, finite sample standard errors for the estimated loadings, unique variances, and factor correlations under the data factor models can be obtained via a bootstrap procedure (e.g., Zhang, 2014).

As a byproduct of our analysis of factor indeterminacy, we provide a mathematical explanation of the empirically observed finding of Schönemann and Wang (1972) and Grice (2001) that factors with less explained (common) variance tend to have smaller minimal correlations. Moreover, together with similar but partial observations by Bartholomew (1981) and Skrondal and Rabe-Hesketh (2004), our analysis explicitly bridges the gap that some observe between the empirical Bayes approach in which factor scores are treated as random, and the fixed factor model in which factor scores are parameters to be estimated.

The paper is organized as follows. In Section 2, we describe the factor models formally. In Section 3 we provide a brief discussion of factor indeterminacy, including some results for imperfect fit. In Section 4 we formulate the assumptions of the data factor model of De Leeuw (2004) (model I) and our constrained version (model II), derive algorithms for the simultaneous estimation of their parameters, and give expressions for the determinate and indeterminate parts of the estimates of the factor scores and the unique parts. Section 5 contains a simulation study in which we compare the performance of the algorithms for models I and II to the existing MINRES method for the random factor model. In Section 6 we fit our model II to a dataset in the literature and demonstrate its practical merits. Finally, Section 7 contains a discussion of our findings.

2. Factor model descriptions

We use the following notation. We write scalars, column vectors, and matrices as z, $\mathbf{z}$, and $\mathbf{Z}$ respectively. The size of a $p \times q$ matrix $\mathbf{Z}$ is specified as $\mathbf{Z} \in \mathbb{R}^{p\times q}$. The transpose is denoted as $\mathbf{z}^T$, the inverse is denoted as $\mathbf{Z}^{-1}$, and we use $\mathbf{Z}^{-T} = (\mathbf{Z}^T)^{-1} = (\mathbf{Z}^{-1})^T$. The $p\times p$ identity matrix is denoted by $\mathbf{I}_p$, a zero matrix is denoted by $\mathbf{O}$, a zero vector is denoted by $\mathbf{0}$,