Computational Statistics and Data Analysis 94 (2016) 265–274

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Fixed factor analysis with clustered factor score constraint

Kohei Uno *, Hironori Satomura, Kohei Adachi

Graduate School of Human Sciences, Osaka University, 1-2 Yamadaoka, Suita, Osaka 565-0871, Japan

ARTICLE INFO

Article history:
Received 25 January 2015
Received in revised form 14 April 2015
Accepted 23 August 2015
Available online 3 September 2015

Keywords:
Fixed factor model
Clustered factor scores
Reduced K-means
Factorial K-means

ABSTRACT

In the fixed factor model for factor analysis (FA), common factor scores are treated as fixed parameters. However, they cannot be estimated jointly with the other parameters, since the maximum likelihood (ML) for the model diverges to infinity. In order to avoid the divergence so that all parameters can be jointly estimated, we propose a constrained fixed factor model. Here, observations are classified into clusters, with each cluster characterized by an equivalent factor score. The ML procedure with the proposed model is named *fixed clustered FA* (FCFA). An iterative algorithm for FCFA is developed, which provides the ML estimates of the factor loadings, unique variances, the classification of observations into clusters, and the cluster factor scores. This FCFA can be viewed as the FA version of Reduced K-means analysis (RKM), in which the principal components are extracted while clustering observations. We compare FCFA, RKM, and a related procedure called Factorial K-means analysis (FKM). We also provide real data examples, which show that FCFA outperforms RKM and FKM in terms of classification accuracy. This result is attributed to the unique variances in FCFA. In other words, the error variances are allowed to be unique to the corresponding variables.

1. Introduction

For an n-individual $\times$ p-variable column-centered data matrix $\mathbf{X}$, the factor analysis(FA) model can be expressed as

$$\mathbf{X} = \mathbf{F}\boldsymbol{\Lambda}' + \mathbf{E} \tag{1}$$

(Mulaik, 2010). Here, $\mathbf{F}$ is the $n \times m$ matrix containing factor scores, Λ is the $p \times m$ factor loadings matrix of full column rank, and each row of $\mathbf{E}$ is a realization of a $1 \times p$ random error vector $\mathbf{e}'$ random error vector $\mathbf{e}'$, with m the number of factors and $m < p < n$. When the distribution of $\mathbf{e}'$ is assumed, it is typically the multivariate normal distribution, with the $p \times 1$ zero vector $\mathbf{0}_p$ as the mean vector:

$$\mathbf{e} \sim N(\mathbf{0}_p, \boldsymbol{\Psi}). \tag{2}$$

Here, the covariance matrix $\boldsymbol{\Psi}$ is a diagonal matrix, with

$$\boldsymbol{\Psi} = \mathrm{diag}\{\psi_1, \ldots, \psi_p\}. \tag{3}$$

The diagonal elements of $\boldsymbol{\Psi}$, $\psi_1, \ldots, \psi_p$, are called unique variances. The parameter matrices to be estimated in FA are $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$. On the other hand, the factor scores in $\mathbf{F}$ are ordinality treated as random variables. In contrast, an FA model in which $\mathbf{F}$ is regarded as a fixed parameter matrix was introduced by Lawley (1942) (see also Young, 1941). This model is called the

* Corresponding author.
E-mail address: uno@bm.hus.osaka-u.ac.jp (K. Uno).

http://dx.doi.org/10.1016/j.csda.2015.08.010
0167-9473/

fixed factor model (e.g., McDonald, 1979; Unkel and Trendafilov, 2010). In this paper, we focus on the fixed factor model. The columns of the factor score matrix $\mathbf{F}$ are centered and orthonormal, with

$$\mathbf{1}_n'\mathbf{F} = \mathbf{0}_m, \tag{4}$$

$$\frac{1}{n}\mathbf{F}'\mathbf{F} = \mathbf{I}_m, \tag{5}$$

in the fixed factor model, where $\mathbf{1}_n$ denotes the $n \times 1$ vector of ones and $\mathbf{I}_m$ is the $m \times m$ identity matrix. The parameter matrices to be estimated in the model are $\mathbf{F}$, $\boldsymbol{\Lambda}$, and $\boldsymbol{\Psi}$. However, it is known that they cannot be jointly estimated (Anderson and Rubin, 1956), as explained in the next section. In this paper, we constrain $\mathbf{F} = [\mathbf{f}_1, \ldots, \mathbf{f}_n]'$ so that the joint estimation is feasible, while $\mathbf{f}_i$ indicates the factor score vector for individual i.

The constraint we consider is that the n individuals in $\mathbf{F}$ are classified into a small number of clusters and the individuals allocated to the same cluster have an equivalent score vector. This constrained fixed factor model underlies a proposed maximum likelihood (ML) procedure, which we refer to as fixed clustered factor analysis (FCFA). The formulation of FCFA is provided in Section 2. We also explain why a joint estimation is possible in FCFA. In Sections 3 and 4, we present the algorithm for FCFA and assess the proposed model using a simulation study.

FCFA is useful for [1] finding the factors underlying variables and [2] clustering individuals, since the model achieves these two purposes simultaneously. The same result can be achieved using two existing procedures, Reduced K-means analysis (RKM) of De Soete and Carroll (1994) and Factorial K-means analysis (FKM) of Vichi and Kiers (2001). However, these analyses are based on principal component analysis rather than FA. Thus, the term factors is replaced by components. We discuss how FCFA is related to RKM and FKM in Section 5. RKM and FKM are least squares procedures, in contrast to the ML-based FCFA. However, we show that the ML-version of RKM can be regarded as a special case of FCFA, with Eq. (3) constrained to be proportional to $\mathbf{I}_p$. In Section 6, we show that FCFA outperforms RKM and FKM in terms of the accuracy of clustering when successively performing [1] and [2].

2. Fixed clustered factor analysis

We begin this section by explaining why the parameters in the original fixed factor model cannot all be jointly estimated. Then, we formulate the proposed FCFA model. Lastly, we describe how joint estimation is possible in FCFA, with the exception of one particular case.

Using (1)–(3), the log-likelihood for the fixed factor model is written as

$$\begin{aligned} LL(\mathbf{F}, \boldsymbol{\Lambda}, \boldsymbol{\Psi}) &\propto -n\sum_{j=1}^{p}\log\psi_j - \mathrm{tr}(\mathbf{X} - \mathbf{F}\boldsymbol{\Lambda}')\boldsymbol{\Psi}^{-1}(\mathbf{X} - \mathbf{F}\boldsymbol{\Lambda}')' \\ &= -n\sum_{j=1}^{p}\frac{1}{\psi_j}\|\mathbf{x}_j - \mathbf{F}\boldsymbol{\lambda}_j\|^2, \end{aligned} \tag{6}$$

where $\mathbf{x}_j$ denotes the jth column of $\mathbf{X}$, $\boldsymbol{\lambda}_j'$ denotes the jth row of $\boldsymbol{\Lambda}$, and ψ_j denotes the jth diagonal element of $\boldsymbol{\Psi}$. However, no maximum likelihood estimate (MLE) exists for (6), which is proved as follows. The estimate of the unique variance must satisfy

$$\psi_j = \frac{1}{n}\|\mathbf{x}_j - \mathbf{F}\boldsymbol{\lambda}_j\|^2. \tag{7}$$

However, this becomes zero, causing (6) to diverge toward infinity as $\mathbf{F}\boldsymbol{\lambda}_j \to \mathbf{x}_j$ so that $\mathbf{x}_j = \mathbf{F}\boldsymbol{\lambda}_j$ (Anderson and Rubin, 1956). For example, $\boldsymbol{\lambda}_j$ can be filled with zeros, with the exception of the jth element taking $s \neq 0$, and the jth row of $\mathbf{F}$ can be $s^{-1}\mathbf{x}_j$, which leads to (7) being equal to zero.

In the proposed FCFA, $\mathbf{F}$ is constrained so that (6) can be maximized to give the MLE of $\mathbf{F}$, $\boldsymbol{\Lambda}$, and $\boldsymbol{\Psi}$. The constraint is that the individuals in $\mathbf{F}$ (i.e., the n rows $\mathbf{f}_1', \ldots, \mathbf{f}_n'$), are classified into K clusters, with $K < n$. This constraint is formally expressed as

$$\mathbf{F} = \mathbf{GC}. \tag{8}$$

Here, $\mathbf{G} = (g_{ik})$ is the n-individual $\times$ K-cluster membership matrix, with $g_{ik} = 1$ if individual i belongs to cluster k, and $g_{ik} = 0$ otherwise:

$$g_{ik} = 0 \text{ or } 1, \mathbf{G}\mathbf{1}_K, \quad \text{and} \quad \mathrm{rank}(\mathbf{G}) = K, \tag{9}$$

where $\mathrm{rank}(\mathbf{G})$ denotes the rank of $\mathbf{G}$. Being equal to K implies that every cluster has at least one individual. On the other hand, $\mathbf{C} = [\mathbf{c}_1, \ldots, \mathbf{c}_r]'$ is a $K \times m$ matrix, with the kth row, $\mathbf{c}_k$, denoting the score vector of cluster k. That is, the factor scores of the individuals belonging to cluster k are constrained to equal $\mathbf{c}_k$. This implies that each row vector of $\mathbf{F} = \mathbf{GC}$ is restricted to one of the K vectors, $\mathbf{c}_1', \ldots, \mathbf{c}_K'$.