Computational Statistics and Data Analysis 72 (2014) 205–218

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Automated learning of factor analysis with complete and incomplete data

Jianhua Zhao, Lei Shi *

School of Statistics and Mathematics, Yunnan University of Finance and Economics, Kunming, 650221, China

HIGHLIGHTS

- We propose a novel algorithm for learning factor analysis with complete and incomplete data.
- The algorithm is able to determine the number of factors in an automated manner.
- The algorithm is as effective as the previous two-stage procedure.
- However, the algorithm is much more computationally efficient.

ARTICLE INFO

Article history:
Received 20 April 2013
Received in revised form 19 September 2013
Accepted 11 November 2013
Available online 20 November 2013

Keywords:
Factor analysis
Model selection
Maximum likelihood
Incomplete data
CM
EM

ABSTRACT

In the application of the popular maximum likelihood method to factor analysis, the number of factors is commonly determined through a two-stage procedure, in which stage 1 performs parameter estimation for a set of candidate models and then stage 2 chooses the best according to certain model selection criterion. Usually, to obtain satisfactory performance, a large set of candidates is used and this procedure suffers a heavy computational burden. To overcome this problem, a novel one-stage algorithm is proposed in which parameter estimation and model selection are integrated in a single algorithm. This is obtained by maximizing the criterion with respect to model parameters and the number of factors jointly, rather than separately. The proposed algorithm is then extended to accommodate incomplete data. Experiments on a number of complete/incomplete synthetic and real data reveal that the proposed algorithm is as effective as the existing two-stage procedure while being much more computationally efficient, particularly for incomplete data.

1. Introduction

Factor analysis (FA) is a commonly used multivariate analysis technique that identifies the common characteristics among a set of variables. The parameter estimation can be easily performed by means of the maximum likelihood (ML) method via the popular expectation maximization (EM)-like algorithm (Dempster et al., 1977; Meng and Rubin, 1993; Meng and van Dyk, 1997). In addition, in density modeling for high-dimensional data, the covariance structure of FA offers significant advantages over full/diagonal/scalar covariance, because of its capability of providing an appropriate trade-off between overfitting full covariance and underfitting diagonal/scalar covariance (Tipping and Bishop, 1999).

To find the trade-off, namely to determine the number of factors q, a two-stage procedure is commonly adopted, where stage 1 performs parameter estimation for a set of candidate models and stage 2 chooses the best according to a model selection criterion. The most popular criteria for this purpose are Akaike's information criterion (AIC) (Akaike, 1987) and the

* Corresponding author. Tel.: +86 0871 65113808; fax: +86 0871 65113808.
E-mail addresses: jhzhao.ynu@gmail.com, jhzhao1@139.com (J. Zhao), shi_lei65@hotmail.com (L. Shi).

0167-9473/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.csda.2013.11.008

Bayesian information criterion (BIC) (Schwarz, 1978). Obviously, this two-stage procedure has to enumerate all candidate models over the set of $\{q\}$. Usually, to achieve satisfactory performance, a large set of values of q is used and thus this procedure can be very time consuming.

To overcome this problem, we propose in this paper a novel one-stage algorithm, in which we do not use a criterion to choose one from a set of candidate models; rather, we integrate the determination of the number of factors into parameter estimation and thus this one-stage algorithm is able to significantly alleviate the computational burden suffered by the two-stage procedure. We call our algorithm the automated learning algorithm of FA (AFA, in short). Specifically, the AFA consists of two conditional maximization (CM)-steps: CM-step 1 maximizes the model selection criterion with respect to (w.r.t.) the number of factors and factor loadings jointly while keeping uniqueness fixed; CM-step 2 maximizes the criterion w.r.t. uniqueness while keeping factor loadings fixed.

In the presence of missing data, Song and Belin (2008) have shown that the AIC and BIC can still be used to determine the number of factors reliably. However, the implementation has to be drawn on the two-stage procedure and, as will be seen in Section 5, its computational burden is typically much heavier than that in the complete data case, due to the inclusion of more missing information. Therefore, we further extend the AFA to accommodate incomplete data by means of the standard methodology in Little and Rubin (1987) and obtain an automated learning algorithm of FA with incomplete data.

Note that the likelihood ratio test (LRT) is also another popular choice for determining the number of factors (Lawley and Maxwell, 1971). However, to the best of our knowledge, there does not exist a corresponding 'automated' version of the LRT just like the AFA for the criterion-based FA in this paper. Thus, the LRT must be run many times for different values of q while the criterion-based FA now can be run only once with the AFA.

The remainder of this paper is organized as follows. In Section 2, we review the conditional maximization (CM) algorithm for fitting FA (Zhao et al., 2008; Yu and Zhao, 2013). Based on this, we propose the AFA for complete data in Section 3 and extend it to accommodate incomplete data in Section 4. We conduct an empirical study to compare the AFA with the existing two-stage procedure in Section 5. We end the paper with some concluding remarks in Section 6.

2. Factor analysis (FA) and a conditional maximization (CM) algorithm

2.1. Factor analysis (FA) model

Suppose that the d-dimensional data vector $\mathbf{x}$ follows a q-factor model:

$$\begin{cases} \mathbf{x} = \mathbf{A}\mathbf{y} + \boldsymbol{\mu} + \boldsymbol{\epsilon}, \\ \mathbf{y} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi}), \end{cases} \tag{1}$$

where $\boldsymbol{\mu}$ is a d-dimensional mean vector, $\mathbf{A}$ is a $d \times q$ factor loading matrix, $\mathbf{y}$ is a q-dimensional latent factor vector, $\boldsymbol{\Psi} = \mathrm{diag}\{\psi_1, \psi_2, \ldots, \psi_d\}$ is a positive diagonal matrix, ψ_i is the so-called uniqueness of variable i in $\mathbf{x}$ and $\mathbf{I}$ denotes an identity matrix of suitable dimension.

If $\mathbf{A}$ in (1) is replaced by $\mathbf{AR}$ and $\mathbf{y}$ by $\mathbf{R}'\mathbf{y}$, where $\mathbf{R}$ is an orthogonal matrix, the FA model is invariant and hence the estimate of $\mathbf{A}$ can only be determined up to a rotation, from which we have that the number of free parameters in the FA model is $\mathcal{D}(q) = d(q+2) - q(q-1)/2$ (Lawley and Maxwell, 1971).

Under model (1),

$$\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{2}$$

where $\boldsymbol{\Sigma} = \mathbf{A}\mathbf{A}' + \boldsymbol{\Psi}$. To avoid over-parameterization, the number of degrees of freedom in $\boldsymbol{\Sigma}$, $d(q+1) - q(q-1)/2$, should not exceed that of a full $d \times d$ covariance matrix, $d(d+1)/2$ (Beal, 2003), which yields

$$q_{\max} \le d + \frac{1}{2}(1 - \sqrt{1 + 8d}). \tag{3}$$

Given a set of i.i.d. observations $\mathbf{X} = \{\mathbf{x}_n\}_{n=1}^N$, let

$$\bar{\mathbf{x}} = \frac{1}{N}\sum_{i=1}^N \mathbf{x}_i, \quad \text{and} \quad \mathbf{S} = \frac{1}{N}\sum_{i=1}^N, \quad (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})' \tag{4}$$

be the sample mean vector and sample covariance matrix of $\mathbf{x}$, respectively. The global maximum likelihood estimator (MLE) of $\boldsymbol{\mu}$ is trivially the sample mean $\bar{\mathbf{x}}$ and thus the MLE of $\boldsymbol{\theta} = (\mathbf{A}, \boldsymbol{\Psi})$ can be obtained by maximizing the log likelihood

$$\mathcal{L}(\boldsymbol{\theta}) = -\frac{N}{2}\{\log|\boldsymbol{\Sigma}| + \mathrm{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{S})\}. \tag{5}$$

There exist many algorithms to maximize $\mathcal{L}$. For example, the expectation maximization (EM) (Rubin and Thayer, 1982), the quasi Newton–Raphson algorithm (Jöreskog, 1967), the expectation CM either (ECME) algorithm (Liu and Rubin, 1998), the parameter-expanded EM (PX-EM) (Liu, 1994), etc. Since these algorithms have to either include latent factors as missing data that yield slow convergence or resort to numerical optimization methods that lack the simplicity and stability of EM,