Computational Statistics and Data Analysis 70 (2014) 61–66

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

# Factor analysis parameter estimation from incomplete data

W.J.J. Roberts *

*Opera Solutions, 10 Exchange Place, Jersey City, NJ 07302, United States*

ARTICLE INFO

*Article history:*
Received 14 November 2012
Received in revised form 30 August 2013
Accepted 30 August 2013
Available online 19 September 2013

*Keywords:*
Recommendation
Expectation–maximization algorithm
Conditional mean

ABSTRACT

An expectation–maximization (EM) algorithm for factor analysis parameter estimation when observations are missing is developed. In contrast to existing EM algorithms for this problem, the algorithm here is developed assuming the missing observations are not part of the complete data in the EM formulation. The resulting algorithm provides increased computational efficiency through sparse matrix operations. The algorithm is demonstrated on two sparse, high-dimensional data sets that are prohibitively large for existing algorithms: the Netflix movie recommendation data set and the Yahoo! musical item data set. The resulting factor models are applied to predict missing values using conditional mean estimation, achieving root mean square errors of 0.9001 and 24.08 on the Netflix and Yahoo! data sets, respectively.



## 1. Introduction

In this paper, we consider factor analysis parameter estimation from incomplete data. Even with complete data, explicit maximum likelihood factor analysis parameter estimation is generally not possible. Expectation–maximization (EM) algorithms developed by Rubin and Thayer (1982), Little and Rubin (2002) and Liu and Rubin (1998) are in widespread use (see, e.g., Zhao et al., 2008, and references therein). When data is incomplete, EM algorithms developed by Jamshidian (1997), Liu and Rubin (1998), and Little and Rubin (2002, p. 235) can be applied. These algorithms require non-sparse matrix operations on the full dimensionality of the data.

Here we develop an EM approach for estimation with incomplete data that has substantial computation advantages compared to existing EM approaches for incomplete data estimation. Existing EM approaches have been developed assuming missing observations are part of the complete data within the EM formulation. The choice of what constitutes complete data within an EM formulation is arbitrary and different choices generally lead to different algorithms. The algorithm here, developed assuming missing observations are not part of the EM complete data, consists of equations involving sparse matrices. The sparsity is straightforwardly exploited to yield substantial computational benefits, particularly for sparse, high-dimensional data.

The algorithm here was applied to factor analysis estimation using two sparse, high-dimensional data sets: the Yahoo! musical preference data set, (see, e.g., Jahrer and Töscher, 2011), and the Netflix movie recommendation data set, (see Bennett and Lanning, 2007). The smaller Netflix data set contains ratings of over 17,000 movies from over 400,000 users. Users rated on average less than 220 movies each. Applying any of the existing factor analysis EM algorithms to this data would require matrix operations on vectors of a dimension over 17,000. The required computation for this data set, and for the larger Yahoo! data set, would be too high for the computing resources available to this study.

* Tel.: +1 201 744 3069.
*E-mail address:* wroberts@operasolutions.com.

0167-9473/$ – see front matter 
http://dx.doi.org/10.1016/j.csda.2013.08.018

## 2. Model specification

Let $Z_t$ denote a $k$-dimensional Gaussian random vector representing observations from the $t$th individual, $t = 1, 2, \ldots, n$. In the exploratory factor analysis model, $Z_t$ is governed by

$$Z_t = \lambda X_t + \mu + W_t \tag{1}$$

where $\mu$ is a $k$-dimensional deterministic mean vector, $\lambda$ is a $k \times p$ deterministic matrix referred to as the coefficient matrix, $X_t$ is a $p$-dimensional Gaussian random vector constituting the latent factors, and $W_t$ is a $k$-dimensional Gaussian random vector. Let $\mathcal{N}(\mu, R)$ denote the Gaussian distribution with mean $\mu$ and covariance $R$. We assume that $\{X_t\}$ is an iid Gaussian process with $X_t \sim \mathcal{N}(0, I_p)$, where $I_p$ is the $p \times p$ identity matrix. We assume that $\{W_t\}$ is an iid Gaussian process with $W_t \sim \mathcal{N}(0, \Sigma)$ where $\Sigma$ is a $k \times k$ diagonal matrix. We assume that $X_t$ and $W_s$ are independent for all $t$ and $s$. Hence $\{Z_t\}$ is an iid process with $Z_t \sim \mathcal{N}(\mu, R)$ where $R = \lambda\lambda' + \Sigma$. We denote the parameter of the factor analysis model as $\phi = \{\mu, \lambda, \Sigma\}$.

With these assumptions, the model (1) is un-identifiable, (see Anderson, 1984, Section 14.2.3). The un-identifiability here is of a specific type, known as spherical un-identifiability, that is generally tolerable in applications. Less restrictive assumptions on $X_t$ and $W_t$ can lead to problematic types of un-identifiability. The model can be made identifiable with additional constraints. The resulting model, however, is generally less tractable than the model considered here. If the covariance of $X_t$ is not diagonal or $\lambda$ has elements that are known to be zero the resulting model is sometimes referred to as confirmatory factor analysis, (see Rubin and Thayer, 1982).

Assume now that for each $t$ only $0 < k_t \leq k$ elements of $Z_t$ are observed. Let $Y_t$ denote a $k_t$-dimensional sub-vector of $Z_t$ representing the observed values of $Z_t$. Let $H_t$ be a deterministic $k_t \times k$ sub-matrix of $I_k$ where the rows of $I_k$ corresponding to the indices of the missing ratings of the $t$th user have been deleted. Thus $Y_t = H_t Z_t$ and

$$Y_t = H_t\lambda X_t + H_t\mu + H_t W_t. \tag{2}$$

For notational convenience we write $\mu_t = H_t\mu$, $R_t = H_t R H_t'$, $\Sigma_t = H_t \Sigma H_t'$ and $\lambda_t = H_t\lambda$. Thus $Y_t \sim \mathcal{N}(\mu_t, R_t)$ where $R_t = \Sigma_t + \lambda_t\lambda_t'$. Note $\mu_t$ and $R_t$ are a sub-vector and a principal sub-matrix of $\mu$ and $R$ respectively. If $R$ is a symmetric and positive semi-definite (psd) matrix then all of its principal sub-matrices are also symmetric and psd (Golub and VanLoan, 1994, Corollary 4.2.2). Thus all such $R_t$ are valid covariance matrices.

Assumptions on the distribution of indices of missing values of $Z_t$ are discussed in detail by Little and Rubin (2002). If we were to assume these indices were *missing completely at random,* rather than deterministic, equations in the corresponding derivation would generally require conditioning on a now-random $H_t$. The resulting approach would otherwise be the same as that derived here.

Independence of $\{Y_t\}$ follows from the independence of $\{Z_t\}$, but $\{Y_t\}$ are not in general iid. Let $Y^n = \{Y_1, \ldots, Y_n\}$ and let $y^n = \{y_1, \ldots, y_n\}$ denote a realization of $Y^n$. The probability density function (pdf) of $Y^n$ is given by

$$p(y^n; \phi) = \prod_{t=1}^{n} \frac{\exp\left(-(y_t - \mu_t)' R_t^{-1} (y_t - \mu_t)/2\right)}{(2\pi)^{k_t/2} |R_t|^{1/2}}. \tag{3}$$

In the subsequent derivations we will have need for the following conditional pdfs. Let $x_t$ be a realization of $X_t$. The conditional pdf of $Y_t$ given $\{X_t = x_t\}$ is given by

$$p(y_t | X_t = x_t; \phi) = \mathcal{N}(\mu_t + \lambda_t x_t, \Sigma_t). \tag{4}$$

The conditional pdf of $X_t$ given $\{Y_t = y_t\}$ is obtained using well-known formulae for conditional Gaussian distributions (see, e.g., Anderson, 1984, Th. 2.5.1) and is given by

$$p(x_t | Y_t = y_t; \phi) = \mathcal{N}(\hat{X}_t, Q_t) \tag{5}$$

where

$$\begin{aligned} \hat{X}_t &= E\{X_t | Y_t = y_t; \phi\} \\ &= \lambda_t' R_t^{-1} (y_t - \mu_t) \end{aligned} \tag{6}$$

and

$$\begin{aligned} Q_t &= E\{(X_t - \hat{X}_t)(X_t - \hat{X}_t)' | Y_t = y_t; \phi\} \\ &= I_p - \lambda_t' R_t^{-1} \lambda_t. \end{aligned} \tag{7}$$

## 3. Model estimation

We aim for a maximum likelihood estimate $\hat{\phi}$ such that

$$\hat{\phi} = \underset{\phi}{\operatorname{argmax}} \log p(y^n; \phi). \tag{8}$$