Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Journal of Multivariate Analysis 95 (2005) 370 – 384

Journal of Multivariate Analysis

www.elsevier.com/locate/jmva

Mixtures of factor analyzers: an extension with covariates

Ernest Fokoué

Department of Statistics, The Ohio State University, Columbus, OH 43210, USA

Received 7 February 2003

Abstract

This paper examines the analysis of an extended finite mixture of factor analyzers (MFA) where both the continuous latent variable (common factor) and the categorical latent variable (component label) are assumed to be influenced by the effects of fixed observed covariates. A polytomous logistic regression model is used to link the categorical latent variable to its corresponding covariate, while a traditional linear model with normal noise is used to model the effect of the covariate on the continuous latent variable. The proposed model turns out be in various ways an extension of many existing related models, and as such offers the potential to address some of the issues not fully handled by those previous models. A detailed derivation of an EM algorithm is proposed for parameter estimation, and latent variable estimates are obtained as by-products of the overall estimation procedure.

AMS subject classification: 62H99

Keywords: Latent variable; Mixtures of factor analyzers; Covariates; Logistic; EM algorithm; Newton–Raphson; Convergence; Generalised linear model

1. Introduction

A finite mixture of factor analyzers (MFA) is a globally nonlinear latent variable model obtained by combining ingredients from the traditional factor analysis (FA) model with ideas from the analysis of finite mixture of distributions. By modelling a local factor analyzer in each subspace of the heterogeneous input space, the MFA model offers a way to overcome the linear limitation of the FA model. Amongst other things, the structure of the MFA

E-mail address: epf@stat.ohio-state.edu.

0047-259X/$ - see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2004.08.004

model offers the potential to model the density of high-dimensional observations adequately while also allowing both clustering and local dimensionality reduction. Many aspects of the MFA model have recently come under close scrutiny, from both the likelihood-based and the Bayesian perspectives. Fokoué [8] reviews the main ingredients of the EM algorithm used for the maximum likelihood estimation of parameters and the estimation of both expected factor scores and posterior class membership for artificial and real examples. Fokoué and Titterington [9] presents a Bayesian analysis of the MFA model, with a treatment that based estimation and inference on the stochastic simulation of the posterior distributions of interest. As noted in [9], the study of the MFA model has indeed received considerable interest in recent years. Refs. [2,5,27] all address the fitting of MFAs or closely related models to psychometrics data using various versions of maximum likelihood estimation (MLE). Ghahramani and Hinton [11] propose an EM algorithm for parameter estimation within the model. Ghahramani and Beal [10] offer a Bayesian treatment of MFA via a variational approximation. Ueda et al. [26] apply their split-and-merge-EM (SMEM) algorithm to the MFA model for such tasks as image compression and handwritten digits recognition. Mclachlan and Peel [18] present a variant of the EM algorithm for a study of the MFA model with application to clustering and density estimation. As an extension and generalisation of two very popular traditional models, namely the finite mixture of distributions and the factor analysis model, the MFA model is very likely to attract more interest from various other scientific communities for a variety of applications.

In its generic formulation, the MFA model focuses solely on the relationship between the manifest variables and the latent variables. This can lead to a neglect of useful information when the latent and/or manifest variables are related to fixed observable covariates. This paper models the effect of covariates on the latent variables.[1] An extension that also takes into account the effect of covariates on the manifest variables is straightforward. The present extension of the MFA model is similar to previous work in latent structures analysis by various authors. From a finite mixture modelling perspective, Aitkin and Wilson [1] used covariates in mixture components as early as 1980. Jansen [13] also considered the use of covariates in mixture modelling and introduced the expression generalised linear finite mixture model to describe his extended model. Following up from [1] and Jansen [13], Thompson et al. [25] have incorporated concomitant information into fixed observed covariates on both the manifest and the latent variables in their assessment of diagnostic criteria for diabetes using a two-component finite mixture model. One of the latest uses of covariates in finite mixtures of distributions can found in [12] who proposed an MCMC implementation in their comprehensive Bayesian analysis of mixtures of regressions. As far as factor analytic models are concerned, Lee and Shi [16] have studied an extension of the structural equation model (SEM) by allowing fixed observed covariates on both the manifest and the latent variables, and have used a Bayesian sampling approach for inference and estimation. Muthen and Shedden [19] used fixed covariates in an extension of a finite

[1] Example: In the application of spatial statistics to disease mapping, it is natural to use mixtures of Poisson distributions to capture the heterogeneity of the distribution of interest. However, in such applications, it would be unrealistic not to take into the spatial dependencies in the estimation of the mixing proportions. It is precisely for this reason that extensions of mixture models like ours that allow covariates to drive the mixing proportions are relevant.

mixture model with mixture outcomes. Finally, Sammel et al. [21] also found it useful to incorporate fixed covariates in their study of latent variable models for mixed discrete and continuous outcomes. The use of fixed observed covariates in the MFA model therefore seems to be justified by such great practical interest. The extended model makes it possible to study, not just the relationship between the manifest and the latent variables, but also the influence of external fixed observed covariates on the latent variables. In the first part of this paper, we give a brief review of some key ingredients of the MFA model needed in this context. We then present the mechanisms by which the covariates are incorporated into the model, after which we give a description of how the EM algorithm is derived for the extended MFA model together with some expressions used in the iterative EM process. The last part is dedicated to simulations and concluding remarks.

2. The mixture of factor analyzers model

Let $X \in \mathbb{R}^p$ be a p-dimensional observed vector of continuous attributes. We assume that X comes from a nonhomogeneous population with k non-overlapping subgroups, and we define $Y \in \{1, \dots, k\}$ to be the random categorical variable that models the group membership. We further assume that X has a simpler structure contained in some continuous latent space of dimension q,[2] where $q < p$, and we define $Z \in \mathbb{R}^q$ to be the intrinsic representation of X. Under the traditional assumptions used for both the orthogonal factor analysis model, $\boldsymbol{p}(\mathbf{x}|Y = j) = \mathcal{N}_p(\mathbf{x}; \boldsymbol{\mu}_j, \Lambda_j \Lambda_j^{\top} + \Sigma)$, where $\mathcal{N}_p$ denotes the p-dimensional normal density, $\Lambda_j \in \mathbb{R}^{p \times q}$ is the matrix of factor loadings, and $\Sigma = \mathsf{diag}(\sigma_1^2, \dots, \sigma_p^2)$ is the diagonal[3] matrix of the specific variances of the X_i's. With $\Pr(Y = j) = \pi_j$, the marginal density of X is given by

$$\boldsymbol{p}(\mathbf{x}) = \sum_{j=1}^{k} \pi_j \mathcal{N}_p(\mathbf{x}; \boldsymbol{\mu}_j, \Lambda_j \Lambda_j^{\top} + \Sigma). \tag{1}$$

It will be very useful in the estimation equations to have a definition of the MFA model in terms of conditional densities. Using all above assumptions,

$$\boldsymbol{p}(\mathbf{x}|\mathbf{z}, Y = j) = \mathcal{N}_p(\mathbf{x}; \boldsymbol{\mu}_j + \Lambda_j \mathbf{z}, \Sigma), \tag{2}$$

where $Z \sim \mathcal{N}_q(0, \mathbf{I}_q)$, with $\mathbf{I}_q$ being the q-dimensional identity matrix.

3. Modelling the effect of covariates

The main motivation for incorporating covariates into the model can be simply stated as follows: *latent variables are related to manifest variables via the mechanism that we have*

[2] Since there are many subgroups in the population, one could assume different intrinsic dimensionalities for each subgroup. In this paper, we focus our attention on the case where q is the same across all groups.

[3] The diagonality of Σ results from assuming conditional independence. Also note that we use the same Σ across all the subgroups of the population. This can be extended easily.

so far modelled with the generic MFA model. However, situations may arise in which those same latent variables are also related to other observables via other mechanisms. As far as the MFA model is concerned, we shall focus in this section on the introduction of two such additional mechanisms: one for the continuous latent variable $\boldsymbol{Z}$ and the other for the categorical latent variable $\boldsymbol{Y}$. Throughout this paper, we shall assume that k and q are known and fixed.

We first assume that each continuous latent variable Z_i is related to a fixed observed covariate $\boldsymbol{w}_i \in \mathbb{R}^r$ through the multivariate linear regression model. The density of Z_i is therefore given by

$$\boldsymbol{p}(\boldsymbol{z}_i|\boldsymbol{w}_i) = \mathcal{N}_q(\boldsymbol{z}_i; \boldsymbol{\Phi}\boldsymbol{w}_i, \boldsymbol{\Psi}), \tag{3}$$

where $\boldsymbol{\Phi}$ is the $q \times r$ matrix of regression parameters, and $\boldsymbol{\Psi} \in \mathbb{R}^{q\times q}$ is the covariance matrix of the noise term. As earlier, we focus on the orthogonal[4] factor structure, and we therefore assume $\boldsymbol{\Psi}$ to be diagonal, that is $\boldsymbol{\Psi} = \mathsf{diag}(\psi_1, \ldots, \psi_q)$. Moreover, since the estimation equations in factor analysis are invariant with respect to scale changes in the factors, as we explained in [9], we retain only the simplest covariance matrix for $\boldsymbol{z}_i$, that is $\boldsymbol{\Psi} = \mathbf{I}_q$. Thus, each $\boldsymbol{Z}_i$ has a multivariate Gaussian distribution, $Z_i \sim \mathcal{N}_q(\boldsymbol{\Phi}\boldsymbol{w}_i, \mathbf{I}_q)$. Essentially, the change brought by the covariate is that the factor score now has a *nonzero* mean, as opposed to the zero mean assumption used for the generic MFA model.[5] We only consider the case of identical $\boldsymbol{\Phi}$ and $\boldsymbol{\Psi} = \mathbf{I}_q$.

We also assume that the categorical latent variable Y is subject to the influence of a fixed observed covariate, $\boldsymbol{u}$, say. Since Y takes its values from $\{1, \ldots, k\}$, a good candidate for dealing with this is the widely used polytomous logistic regression model. Given a vector $\boldsymbol{u} \in \mathbb{R}^s$ of covariates, the prior classification probabilities are defined through the logit model as follows:

$$\log\left[\frac{\Pr(\boldsymbol{Y} = j|\boldsymbol{u})}{\Pr(\boldsymbol{Y} = k|\boldsymbol{u})}\right] = \boldsymbol{\phi}_j^{\mathsf{T}}\boldsymbol{u} = \boldsymbol{u}^{\mathsf{T}}\boldsymbol{\phi}_j \quad \text{for} \quad j = 1, \ldots, k-1, \tag{4}$$

where $\boldsymbol{\phi}_j^{\mathsf{T}} = (\phi_{0j}, \phi_{1j}, \ldots, \phi_{s-1,j}) \in \mathbb{R}^s$, for $j = 1, \ldots, k-1$ and $\boldsymbol{u}^{\mathsf{T}} = (1, u_1, \ldots, u_{s-1}) \in \mathbb{R}^s$. It is easy to show from (4) that the prior classification probabilities are given by

$$\Pr(\boldsymbol{Y} = j|\boldsymbol{u}) = \frac{\exp(\boldsymbol{u}^{\mathsf{T}}\boldsymbol{\phi}_j)}{1 + \sum_{j'=1}^{k-1}\exp(\boldsymbol{u}^{\mathsf{T}}\boldsymbol{\phi}_{j'})} \quad \text{for} \quad j = 1, \ldots, k, \tag{5}$$

where we set $\boldsymbol{\phi}_k = 0$ for identifiability.

For simplicity and convenience, we define $\pi_{ij} = \pi_{ij}(\boldsymbol{u}_i, \boldsymbol{\phi}_j) = \Pr(Y_i = j|\boldsymbol{u}_i)$ for $j = 1, \ldots, k$ and $i = 1, \ldots, n$. For economy of notational space, we shall omit the explicit

[4] We assume factor scores to be uncorrelated.

[5] It is possible to imagine a more general extension in which there is a different $\boldsymbol{\Phi}_j$ for each component j of the mixture, and where $\boldsymbol{\Psi}$ is a full variance–covariance matrix reflecting the fact that factors are allowed to be correlated.

mention of covariates and parameters in many of our expressions of probability densities and expectations, unless a need for clarity requires it. For instance, we shall simply write $[\mathbf{x}_i | Y_i = j]$ instead of $[\mathbf{x}_i | \boldsymbol{Y}_i = j, \boldsymbol{w}_i, \boldsymbol{\theta}]$, and $\Pr(Y = j)$ instead of $\Pr(Y = j | \boldsymbol{u}, \boldsymbol{\theta})$.

It is straightforward to verify that

$$\boldsymbol{p}(\mathbf{x}_i | Y_i = j) = \mathcal{N}_p(\mathbf{x}_i; \mu_j + \boldsymbol{\Lambda}_j \boldsymbol{\Phi} \boldsymbol{w}_i, \boldsymbol{\Lambda}_j \boldsymbol{\Lambda}_j^\top + \boldsymbol{\Sigma}) \tag{6}$$

and that the corresponding marginal density of X_i is now

$$\boldsymbol{p}(\mathbf{x}_i) = \sum_{j=1}^{k} \pi_{ij} \mathcal{N}_p(\mathbf{x}_i; \mu_j + \boldsymbol{\Lambda}_j \boldsymbol{\Phi} \boldsymbol{w}_i, \boldsymbol{\Lambda}_j \boldsymbol{\Lambda}_j^\top + \boldsymbol{\Sigma}). \tag{7}$$

As we shall see later, it turns out to be more convenient to reformulate our model here as a multivariate generalised linear model (GLM) for multicategorical responses. More specifically, we now consider the $(k-1)$-dimensional vector of indicator variables $Y_i = (Y_{i1}, \ldots, Y_{i,k-1})^\top$, with $Y_{ij} = 1$ if $Y_i = j$ and 0 otherwise.

We define $\boldsymbol{U}_i \in \mathbb{R}^{(k-1)s \times (k-1)s}$, $\boldsymbol{\phi} \in \mathbb{R}^{(k-1)s}$, and $\boldsymbol{\pi}_i \in \mathbb{R}^{k-1}$ as follows:

$$\boldsymbol{U}_i = \begin{bmatrix} \boldsymbol{u}_i^\top & & & \\ & \boldsymbol{u}_i^\top & & \\ & & \ddots & \\ & & & \boldsymbol{u}_i^\top \end{bmatrix}, \quad \boldsymbol{\phi} = \begin{bmatrix} \boldsymbol{\phi}_1 \\ \boldsymbol{\phi}_2 \\ \vdots \\ \boldsymbol{\phi}_{k-1} \end{bmatrix} \quad \text{and} \quad \boldsymbol{\pi}_i = \begin{bmatrix} \pi_{i1} \\ \pi_{i2} \\ \vdots \\ \pi_{i,k-1} \end{bmatrix}. \tag{8}$$

From the above definitions, the *systematic component* of our GLM for a given covariate $\boldsymbol{u}_i$ is the vector $\boldsymbol{\eta}_i = \boldsymbol{U}_i \boldsymbol{\phi} = (\boldsymbol{\eta}_{i1}, \ldots, \boldsymbol{\eta}_{i,k-1})^\top$, with $\boldsymbol{\eta}_{ij} = \boldsymbol{u}_i^\top \boldsymbol{\phi}_j$, for $j = 1, \ldots, k-1$. The response function here is a vector-valued function $\boldsymbol{f} = (\boldsymbol{f}_1, \ldots, \boldsymbol{f}_{k-1})$, with

$$\boldsymbol{f}_j(\boldsymbol{\eta}_i) = \frac{\exp(\boldsymbol{\eta}_{ij})}{1 + \sum_{j'=1}^{k-1} \exp(\boldsymbol{\eta}_{ij'})}, \quad j = 1, \ldots, k-1, \tag{9}$$

which allows us to express the $\boldsymbol{\pi}_i$ of Eq. (8) as $\boldsymbol{\pi}_i = \boldsymbol{f}(\boldsymbol{\eta}_i) = \boldsymbol{f}(\boldsymbol{U}_i \boldsymbol{\phi})$. Expressed in terms of the link function of the logit model, we have $\boldsymbol{\eta}_i = \boldsymbol{g}(\boldsymbol{\pi}_i) = \boldsymbol{U}_i \boldsymbol{\phi}$, where $\boldsymbol{g} = (\boldsymbol{g}_1, \ldots, \boldsymbol{g}_{k-1})$ is a vector-valued function such that

$$\boldsymbol{g}_j(\boldsymbol{\pi}_i) = \log\left[\frac{\pi_{ij}}{1 - (\pi_{i1} + \cdots + \pi_{i,k-1})}\right]. \tag{10}$$

The variance–covariance matrix for a given variable $\boldsymbol{y}_i = (Y_{i1}, \ldots, Y_{i,k-1})^\top$ is

$$C_i = C_i(\boldsymbol{\phi}) = \begin{bmatrix} \pi_{i1}(1-\pi_{i1}) & -\pi_{i1}\pi_{i2} & \cdots & -\pi_{i1}\pi_{i,k-1} \\ -\pi_{i2}\pi_{i1} & \pi_{i2}(1-\pi_{i2}) & \cdots & -\pi_{i2}\pi_{i,k-1} \\ \vdots & \vdots & \ddots & \\ -\pi_{i,k-1}\pi_{i1} & \cdots & \cdots & \pi_{i,k-1}(1-\pi_{i,k-1}) \end{bmatrix}. \tag{11}$$

It is easy to verify that $C_i = \mathsf{diag}(\boldsymbol{\pi}_i) - \boldsymbol{\pi}_i \boldsymbol{\pi}_i^\top$. Our complete collection of model parameters is now $\boldsymbol{\theta} = \{\boldsymbol{\phi}, \boldsymbol{\Lambda}, \boldsymbol{\mu}, \Sigma, \Phi\}$ where $\boldsymbol{\phi} = \{\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_{k-1}\}$.

4. Elements of parameter estimation

Modelling the effect of covariates on latent variables can only be fully justified if the estimation of latent scores plays a central (key) role in the statistical analysis being carried out. It is therefore important in this context to concentrate a large amount of effort on addressing the estimation of both posterior expectations of factor scores and posterior classification probabilities. Parameter estimation obviously remains the prime focus, since the other inferential tasks depend on it.

4.1. Conditional posterior expectations for Y and Z

In each component *j* of the mixture, we have the following distribution:

$$\begin{bmatrix} Z \\ X \end{bmatrix} \sim \mathcal{N}_{(q+p)}\left(\begin{bmatrix} \boldsymbol{\Phi w} \\ \mu_j + \Lambda_j \boldsymbol{\Phi w} \end{bmatrix}, \begin{bmatrix} \mathbf{I}_q & (\mathbf{I}_q - \Phi \boldsymbol{w}\boldsymbol{w}^\top \Phi^\top)\Lambda_j^\top \\ \Lambda_j(\mathbf{I}_q - \Phi \boldsymbol{w}\boldsymbol{w}^\top \Phi^\top) & \boldsymbol{\Sigma} + \boldsymbol{\Lambda}_j \boldsymbol{\Lambda}_j^\top \end{bmatrix}\right). \tag{12}$$

It is easy to establish that the conditional density of *Z* given *X* and *Y* is

$$\boldsymbol{p}(\mathbf{z}|\mathbf{x}, Y = j) = \mathcal{N}_q(\mathbf{z}; m_{\mathbf{z}|\mathbf{x},\boldsymbol{y}=j}, C_{\mathbf{z}|\mathbf{x},\boldsymbol{y}=j}),$$

where

$$\begin{aligned} m_{Z|\mathbf{x},\boldsymbol{y}=j} &= \boldsymbol{\Phi w} + (\mathbf{I}_q - \boldsymbol{\Phi w w}^\top \boldsymbol{\Phi}^\top)\boldsymbol{\Lambda}_j^\top \left(\boldsymbol{\Lambda}_j \boldsymbol{\Lambda}_j^\top + \boldsymbol{\Sigma}\right)^{-1} (\mathbf{x} - \mu_j - \Lambda_j \boldsymbol{\Phi w}), \\ C_{Z|\mathbf{x},\boldsymbol{y}=j} &= \mathbf{I}_q - (\mathbf{I}_q - \boldsymbol{\Phi w w}^\top \boldsymbol{\Phi}^\top)\boldsymbol{\Lambda}_j^\top \left(\boldsymbol{\Lambda}_j \boldsymbol{\Lambda}_j^\top + \boldsymbol{\Sigma}\right)^{-1} \boldsymbol{\Lambda}_j (\mathbf{I}_q - \boldsymbol{\Phi w w}^\top \boldsymbol{\Phi}^\top). \end{aligned} \tag{13}$$

Thus, given an observation $\mathbf{x}_i$, a covariate $\boldsymbol{w}_i$, an assumed value y_{ij} of the label of $\mathbf{x}_i$ and a set of parameters $\boldsymbol{\theta}$, an estimate of the expected factor score is given by $\mathbf{E}\left[Z_i|\boldsymbol{w}_i, \mathbf{x}_i, \boldsymbol{y}_i = j\right] = \mathbf{E}\left[Z_i|\cdots\right]$ where

$$\mathbf{E}\left[Z_i|\cdots\right] = \boldsymbol{\Phi w}_i + (\mathbf{I}_q - \boldsymbol{\Phi w}_i \boldsymbol{w}_i^\top \boldsymbol{\Phi}^\top)\boldsymbol{\Lambda}_j^\top \left(\boldsymbol{\Lambda}_j \boldsymbol{\Lambda}_j^\top + \boldsymbol{\Sigma}\right)^{-1} (\mathbf{x}_i - \mu_j - \Lambda_j \boldsymbol{\Phi w}_i). \tag{14}$$

The posterior classification probabilities are now given by

$$\Pr(Y_{ij} = 1|\mathbf{x}_i) = \frac{\pi_{ij}\mathcal{N}_p(\mathbf{x}_i; \mu_j + \boldsymbol{\Lambda}_j \boldsymbol{\Phi w}_i, \boldsymbol{\Lambda}_j \boldsymbol{\Lambda}_j^\top + \boldsymbol{\Sigma})}{\sum_{j'=1}^{k} \pi_{ij'}\mathcal{N}_p(\mathbf{x}_i; \mu_{j'} + \boldsymbol{\Lambda}_{j'} \boldsymbol{\Phi w}_i, \boldsymbol{\Lambda}_{j'} \boldsymbol{\Lambda}_{j'}^\top + \boldsymbol{\Sigma})}, \tag{15}$$

where $\pi_{ij} = \pi_{ij}(\boldsymbol{u}_i, \boldsymbol{\phi}_j) = \Pr(Y_{ij} = 1)$ and $\mathbf{E}\left[Y_{ij}|\mathbf{x}_i\right] = \Pr(Y_{ij} = 1|\mathbf{x}_i)$.

4.2. The expected complete data log likelihood

The above estimates of posterior expected factor scores (14) and posterior classification probabilities (15) presuppose the existence of a set of parameter estimates. In this paper, we only tackle parameter estimation from a likelihood-based perspective via the EM algorithm.

The EM algorithm for this extended MFA model makes extensive use of elements from [8]. In fact, the joint density of all the variables is now

$$\boldsymbol{p}(\mathbf{x}, \boldsymbol{y}, \boldsymbol{z}) = \boldsymbol{p}(\mathbf{x}|\boldsymbol{y}, \boldsymbol{z})\boldsymbol{p}(\boldsymbol{y}|\boldsymbol{u})\boldsymbol{p}(\boldsymbol{z}|\boldsymbol{w}). \tag{16}$$

The complete-data log-likelihood of the model is therefore given by

$$\begin{aligned}\boldsymbol{\ell}(\boldsymbol{\theta}; \mathbf{X}^*) = {} & \sum_{i=1}^{n}\sum_{j=1}^{k} y_{ij} \log \boldsymbol{p}(\mathbf{x}_i|Y_{ij} = 1, \boldsymbol{z}_i) + \sum_{i=1}^{n}\sum_{j=1}^{k} y_{ij} \log \pi_{ij} \\ & + \sum_{i=1}^{n} \log \boldsymbol{p}(\boldsymbol{z}_i|\boldsymbol{w}_i). \end{aligned}\tag{17}$$

The expectation of the complete-data log-likelihood with respect to the joint conditional distribution of Y and Z given $\mathbf{X}$ and $\boldsymbol{\theta}^{(t)}$ is defined as

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \mathbf{E}_{(\boldsymbol{y},\boldsymbol{z})}\left[\boldsymbol{\ell}(\boldsymbol{\theta}, \mathbf{X}^*)|\mathbf{X}, \boldsymbol{\theta}^{(t)}\right] = \int_{\mathcal{H}} \boldsymbol{\ell}(\boldsymbol{\theta}, \mathbf{X}^*)\boldsymbol{p}(\boldsymbol{y}, \mathbf{z}|\mathbf{X}, \boldsymbol{\theta}^{(t)})\, d\boldsymbol{y}\, d\mathbf{z}. \tag{18}$$

In order to perform the exact EM algorithm, we need to construct an analytical expression for $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$. As usual, the iterations of the EM algorithm proceed with the following two steps:

$$\begin{aligned} &E\text{-}step\text{— Find an analytical expression for } Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}),\\ &M\text{-}step\text{— Solve } \boldsymbol{\theta}^{(t+1)} = \underset{\boldsymbol{\theta}}{\mathbf{argmax}}\quad Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}).\end{aligned}$$

Since all our expectations are taken with respect to the joint distribution of $(\boldsymbol{y}, \mathbf{z})$ conditional on $\mathbf{X}$ and $\boldsymbol{\theta}^{(t)}$, we simply use $\mathbf{E}$ instead of $\mathbf{E}_{(\boldsymbol{y},\boldsymbol{z})}$.

Based on the expression of $\boldsymbol{\ell}(\boldsymbol{\theta}, \mathbf{X}^*)$ in Eq. (17), the formation of an analytical expression for $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ in (18) requires analytical expressions for $\mathbf{E}\left[\boldsymbol{z}_i|\mathbf{x}_i, \boldsymbol{\theta}^{(t)}\right]$ and $\mathbf{E}\left[\boldsymbol{z}_i\mathbf{z}_i^{\mathsf{T}}|\mathbf{x}_i, \boldsymbol{\theta}^{(t)}\right]$. Besides, we define $\mathsf{a}_{ij}^{(t)} = \mathbf{E}\left[y_{ij}|\mathbf{x}_i, \boldsymbol{\theta}^{(t)}\right]$, along with $\mathbf{b}_{ij}^{(t)} = \mathbf{E}\left[\boldsymbol{z}_i|y_{ij} = 1, \mathbf{x}_i, \boldsymbol{\theta}^{(t)}\right]$ and $\mathbf{C}_{ij}^{(t)} = \mathbf{E}\left[\boldsymbol{z}_i\boldsymbol{z}_i^{\mathsf{T}}|y_{ij} = 1, \mathbf{x}_i, \boldsymbol{\theta}^{(t)}\right]$.

From the fact that $\mathbf{E}_{(\boldsymbol{y},\boldsymbol{z})}\left[\boldsymbol{z}_i|\mathbf{x}_i, \boldsymbol{\theta}^{(t)}\right] = \mathbf{E}_{\boldsymbol{y}}\left[\mathbf{E}_{\mathbf{z}}\left[\boldsymbol{z}_i|\mathbf{x}_i, \boldsymbol{y}_i, \boldsymbol{\theta}^{(t)}\right]\right]$, we get

$$\mathbf{E}\left[\boldsymbol{z}_i|\mathbf{x}_i, \boldsymbol{\theta}^{(t)}\right] = \sum_{j=1}^{k} \mathsf{a}_{ij}^{(t)}\mathbf{b}_{ij}^{(t)} \quad \text{and} \quad \mathbf{E}\left[\boldsymbol{z}_i\boldsymbol{z}_i^{\mathsf{T}}|\mathbf{x}_i, \boldsymbol{\theta}^{(t)}\right] = \sum_{j=1}^{k} \mathsf{a}_{ij}^{(t)}\mathbf{C}_{ij}^{(t)}. \tag{19}$$

With the above expressions clearly defined, the derivation of $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ turns out to be straightforward, making the E-step an easy one in this case. However, as we shall see later, some of the parameters do not allow direct analytical updating at the M-step. Nevertheless, it is reassuring to know that the Newton–Raphson iteration used to find new updates turns out to behave well, thanks to the good properties of the function of interest.

With the incorporation of fixed observed covariates into our model, we now have to obtain the mixing proportions through their corresponding parameters $\boldsymbol{\phi}_j$. As a function of $\boldsymbol{\phi}$, the function Q can be written as

$$Q(\boldsymbol{\phi}) = \mathbf{E}\left[\sum_{i=1}^{n}\sum_{j=1}^{k} y_{ij}\ \log(\pi_{ij}(\boldsymbol{u}_i, \boldsymbol{\phi}_j))\right] = \sum_{i=1}^{n}\sum_{j=1}^{k} \mathsf{a}_{ij}^{(t)}\ \log(\pi_{ij}(\boldsymbol{u}_i, \boldsymbol{\phi}_j)). \tag{20}$$

Recall that our aim at the M-Step is to find a new $\boldsymbol{\phi}$ that maximises $Q(\boldsymbol{\phi})$ subject to

$$\sum_{j=1}^{k} \pi_{ij} = 1 \quad \text{and} \quad \sum_{j=1}^{k} \mathsf{a}_{ij}^{(t)} = 1. \tag{21}$$

4.3. Estimation of $\boldsymbol{\phi}$ for a 2-component mixture

We first restrict ourselves to a 2-component mixture in order to gain more insights into the estimation of $\boldsymbol{\phi}$. In fact, if we only have two components, then $\boldsymbol{y}$ has a Bernoulli distribution $\mathsf{Ber}(\pi)$, where $\pi = \pi(\boldsymbol{\phi}, \boldsymbol{u})$ is a function of $\boldsymbol{u}$ and $\boldsymbol{\phi}$ defined as follows:

$$\Pr(\boldsymbol{Y}_i = 1|\boldsymbol{u}_i) = \pi_i = \frac{\exp(\boldsymbol{u}_i^{\mathsf{T}}\boldsymbol{\phi})}{1 + \exp(\boldsymbol{u}_i^{\mathsf{T}}\boldsymbol{\phi})}. \tag{22}$$

From (22) and (21), our function Q in this binary case is now

$$Q(\boldsymbol{\phi}) = \sum_{i=1}^{n} \mathsf{a}_i^{(t)} \log(\pi_i) + (1 - \mathsf{a}_i^{(t)}) \log(1 - \pi_i). \tag{23}$$

It is easy to see that $Q(\boldsymbol{\phi})$ is a nonlinear function of $\boldsymbol{\phi}$. On the other hand, it is important to note that the form of $Q(\boldsymbol{\phi})$ does not allow the derivation of a closed-form expression for its maximiser. We use Newton–Raphson iteration to find the maximiser, obtained by solving the equation $\dfrac{\partial Q}{\partial \boldsymbol{\phi}} = 0$.

$$\frac{\partial Q}{\partial \pi_i} = \frac{\mathsf{a}_i^{(t)} - \pi_i}{\pi_i(1 - \pi_i)} \quad \text{and} \quad \frac{\partial \pi_i}{\partial \boldsymbol{\phi}} = \pi_i(1 - \pi_i)\boldsymbol{u}_i. \tag{24}$$

If we use the chain rule $\dfrac{\partial Q}{\partial \boldsymbol{\phi}} = \dfrac{\partial Q}{\partial \pi_i}\dfrac{\partial \pi_i}{\partial \boldsymbol{\phi}}$, it is straightforward to find that

$$\frac{\partial Q}{\partial \boldsymbol{\phi}} = \sum_{i=1}^{n} (\mathsf{a}_i^{(t)} - \pi_i)\boldsymbol{u}_i = F(\boldsymbol{\phi}). \tag{25}$$

The matrix $J(\boldsymbol{\phi})$ of first derivatives of $F(\boldsymbol{\phi})$ in this case is given by

$$J(\boldsymbol{\phi}) = \frac{\partial F}{\partial \boldsymbol{\phi}} = -\sum_{i=1}^{n} \pi_i(1 - \pi_i)\boldsymbol{u}_i\boldsymbol{u}_i^{\mathsf{T}}. \tag{26}$$

With F and J thus defined, the update $\boldsymbol{\phi}^{(t+1)}$ of $\boldsymbol{\phi}$ at iteration $t+1$ of the EM algorithm is obtained by Newton–Raphson iterations with update equation

$$\boldsymbol{\phi}^{\text{new}}(m) := \boldsymbol{\phi}^{\text{new}}(m-1) - J^{-1}(\boldsymbol{\phi}^{\text{new}}(m-1))F(\boldsymbol{\phi}^{\text{new}}(m-1)). \tag{27}$$

At each step of the EM algorithm, (27) is run until a chosen tolerance is reached. It is obviously important to remember, the behaviour of Newton–Raphson iterations, in terms of convergence and stability may depend on the accuracy of initial guesses and the existence of $J^{-1}(\boldsymbol{\phi})$.

According to a standard Newton–Raphson property, (27) achieves local quadratic convergence if its initial values are accurate enough and $J^{-1}(\boldsymbol{\phi}_j)$ exists. In our context, it is easy to see that the matrix $J(\boldsymbol{\phi})$ of first derivatives of $F(\boldsymbol{\phi})$ defined by (26) is negative definite. In fact, since $\boldsymbol{u}_i\boldsymbol{u}_i^\top$ is a positive semi-definite matrix, and the term $\pi_i(1-\pi_i)$ is a positive number, the sum $\sum_{i=1}^n \pi_i(1-\pi_i)\boldsymbol{u}_i\boldsymbol{u}_i^\top$ is therefore a positive definite matrix, and as a result, $J(\boldsymbol{\phi})$ is a negative definite matrix. Finally, with $J(\boldsymbol{\phi})$ is negative definite, $J^{-1}(\boldsymbol{\phi})$ exists, and (27) should therefore require very few iterations to yield the desired updates.

4.4. *Estimation of $\boldsymbol{\phi}$ for a k-component mixture*

If we use the GLM formulation of Section 3, then we can rewrite $Q(\boldsymbol{\phi})$ as

$$Q(\boldsymbol{\phi}) \propto \sum_{i=1}^{n}\left[[\mathsf{a}_i^{(t)}]^\top\boldsymbol{\eta}_i - \mathsf{b}(\boldsymbol{\eta}_i)\right], \tag{28}$$

where $\mathsf{a}_i^{(t)} = (\mathsf{a}_{i1}^{(t)}, \ldots, \mathsf{a}_{ik-1}^{(t)})^\top$ and $\mathsf{b}(\boldsymbol{\eta}_i) = \log(1+\sum_{j=1}^{k-1}\exp(\boldsymbol{\eta}_{ij}))$, so that

$$\frac{\partial\mathsf{b}(\boldsymbol{\eta}_i)}{\partial\boldsymbol{\eta}_i} = \boldsymbol{\pi}_i.$$

It is also easy to show that

$$\frac{\partial\boldsymbol{\pi}_i}{\partial\boldsymbol{\eta}_i} = \frac{\partial^2\mathsf{b}(\boldsymbol{\eta}_i)}{\partial\boldsymbol{\eta}_i^2} = C_i(\boldsymbol{\phi}),$$

where $C_i(\boldsymbol{\phi})$ is as defined in (11). By the chain rule, we have

$$\frac{\partial\boldsymbol{\pi}_i}{\partial\boldsymbol{\phi}} = \frac{\partial\boldsymbol{\pi}_i}{\partial\boldsymbol{\eta}_i}\frac{\partial\boldsymbol{\eta}_i}{\partial\boldsymbol{\phi}} = \frac{\partial\boldsymbol{\pi}_i}{\partial\boldsymbol{\eta}_i}\boldsymbol{U}_i = C_i(\boldsymbol{\phi})\boldsymbol{U}_i.$$

From the above definition of Q in (28), and considering the fact that our logistic link function is a canonical link function, a well-established result in GLM theory [4,7] allows us to easily derive F and J as follows:

$$F(\boldsymbol{\phi}) = \frac{\partial Q}{\partial\boldsymbol{\phi}} = \sum_{i=1}^{n}\boldsymbol{U}_i^\top\left[\mathsf{a}_i^{(t)} - \boldsymbol{\pi}_i\right] \quad\text{and}\quad J(\boldsymbol{\phi}) = \frac{\partial F}{\partial\boldsymbol{\phi}} = -\sum_{i=1}^{n}\boldsymbol{U}_i^\top C_i(\boldsymbol{\phi})\boldsymbol{U}_i. \tag{29}$$

Just as before, it is easy to see that $J(\boldsymbol{\phi})$ as defined by (29) is negative definite, so that $J^{-1}(\boldsymbol{\phi})$ exists, thereby guaranteeing quadratic local convergence to the update $\boldsymbol{\phi}^{(t+1)}$.

4.5. *Estimating the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Lambda}$*

Estimating the means $\mu_j^{(t+1)}$ of the Gaussians is rather straightforward. In fact, as a function of $\boldsymbol{\mu}$, the function Q can be written as

$$Q(\boldsymbol{\mu}) = -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{k}\mathsf{a}_{ij}^{(t)}\left[-2\mathbf{x}_i^{\mathsf{T}}\boldsymbol{\Sigma}^{-\mathbf{1}}\boldsymbol{\mu}_j + 2\boldsymbol{\mu}_j^{\mathsf{T}}\boldsymbol{\Sigma}^{-\mathbf{1}}\boldsymbol{\Lambda}_j\mathbf{b}_{ij}^{(t)} + \boldsymbol{\mu}_j^{\mathsf{T}}\boldsymbol{\Sigma}^{-\mathbf{1}}\boldsymbol{\mu}_j\right].$$

Since $\dfrac{\partial\mu_j^{\mathsf{T}}\boldsymbol{\Sigma}^{-\mathbf{1}}\mu_j}{\partial\mu_j} = 2\boldsymbol{\Sigma}^{-\mathbf{1}}\mu_j$, $\dfrac{\partial\mathbf{x}_i^{\mathsf{T}}\boldsymbol{\Sigma}^{-\mathbf{1}}\mu_j}{\partial\mu_j} = \boldsymbol{\Sigma}^{-\mathbf{1}}\mathbf{x}_i^{\mathsf{T}}$ and $\dfrac{\partial\boldsymbol{\mu}_j^{\mathsf{T}}\boldsymbol{\Sigma}^{-\mathbf{1}}\boldsymbol{\Lambda}_j\mathbf{b}_{ij}^{(t)}}{\partial\mu_j} = \boldsymbol{\Sigma}^{-\mathbf{1}}\boldsymbol{\Lambda}_j\mathbf{b}_{ij}^{(t)}$, it is easy to show that solving $\dfrac{\partial Q(\boldsymbol{\mu})}{\partial\mu_j} = 0$ yields the maximiser of $Q(\boldsymbol{\mu})$ that is given by

$$\boldsymbol{\mu}_j^{(t+1)} = \left[\sum_{i=1}^{n}\mathsf{a}_{ij}^{(t)}\left(\mathbf{x}_i - \boldsymbol{\Lambda}_j^{(t)}\mathbf{b}_{ij}^{(t)}\right)\right]\left[\sum_{i'=1}^{n}\mathsf{a}_{i'j}^{(t)}\right]^{-1}.$$

If we treat Q as a function of $\boldsymbol{\Lambda}$, we can express it as

$$Q(\boldsymbol{\Lambda}) = \sum_{i=1}^{n}\sum_{j=1}^{k}\mathsf{a}_{ij}^{(t)}(\mathbf{x}_i - \mu_j)^{\mathsf{T}}\boldsymbol{\Sigma}^{-\mathbf{1}}\boldsymbol{\Lambda}_j\mathbf{b}_{ij}^{(t)} - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{k}\mathsf{a}_{ij}^{(t)}\mathbf{tr}\left[\boldsymbol{\Lambda}_j^{\mathsf{T}}\boldsymbol{\Sigma}^{-\mathbf{1}}\boldsymbol{\Lambda}_j\mathbf{C}_{ij}^{(t)}\right]$$

and the corresponding partial derivatives with respect to Λ_j are given by

$$\begin{aligned}\frac{\partial(\mathbf{x}_i - \mu_j)^{\mathsf{T}}\boldsymbol{\Sigma}^{-\mathbf{1}}\boldsymbol{\Lambda}_j\mathbf{b}_{ij}^{(t)}}{\partial\Lambda_j} &= \boldsymbol{\Sigma}^{-\mathbf{1}}(\mathbf{x}_i - \mu_j)\left(\mathbf{b}_{ij}^{(t)}\right)^{\mathsf{T}},\\ \frac{\partial\mathbf{tr}(\boldsymbol{\Lambda}_j^{\mathsf{T}}\boldsymbol{\Sigma}^{-\mathbf{1}}\boldsymbol{\Lambda}_j\mathbf{C}_{ij}^{(t)})}{\partial\Lambda_j} &= 2\boldsymbol{\Sigma}^{-\mathbf{1}}\Lambda_j\mathbf{C}_{ij}^{(t)}.\end{aligned}$$

The solution of $\dfrac{\partial Q(\boldsymbol{\Lambda})}{\partial\Lambda_j} = 0$ yields the maximiser of $Q(\boldsymbol{\Lambda})$ which is

$$\boldsymbol{\Lambda}_j^{(t+1)} = \left[\sum_{i=1}^{n}\mathsf{a}_{ij}^{(t)}(\mathbf{x}_i - \boldsymbol{\mu}_j^{(t+1)})\left(\mathbf{b}_{ij}^{(t)}\right)^{\mathsf{T}}\right]\left[\sum_{i'=1}^{n}\mathsf{a}_{i'j}^{(t)}\mathbf{C}_{i'j}^{(t)}\right]^{-1}.$$

4.6. *Estimating the uniquenesses $\boldsymbol{\Sigma}$*

As a function of $\boldsymbol{\Sigma}$, Q can be written as follows:

$$\begin{aligned}Q(\boldsymbol{\Sigma}) = &-\frac{n}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{k}\mathsf{a}_{ij}^{(t)}\mathbf{tr}\left[\boldsymbol{\Sigma}^{-\mathbf{1}}(\mathbf{x}_i - \mu_j)(\mathbf{x}_i - \mu_j)^{\mathsf{T}}\right]\\ &+\sum_{i=1}^{n}\sum_{j=1}^{k}\mathsf{a}_{ij}^{(t)}(\mathbf{x}_i - \mu_j)^{\mathsf{T}}\boldsymbol{\Sigma}^{-\mathbf{1}}\boldsymbol{\Lambda}_j\mathbf{b}_{ij}^{(t)} - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{k}\mathsf{a}_{ij}^{(t)}\mathbf{tr}\left[\boldsymbol{\Lambda}_j^{\mathsf{T}}\boldsymbol{\Sigma}^{-\mathbf{1}}\boldsymbol{\Lambda}_j\mathbf{C}_{ij}^{(t)}\right].\end{aligned}$$

It is easier to derive the partial derivatives of $Q(\mathbf{\Sigma})$ with respect to $\mathbf{\Sigma^{-1}}$, namely

$$\begin{aligned}
\frac{\partial \mathbf{tr}\left[\mathbf{\Sigma^{-1}}(\mathbf{x}_i-\mu_j)(\mathbf{x}_i-\mu_j)^{\mathsf{T}}\right]}{\partial \mathbf{\Sigma^{-1}}} &= (\mathbf{x}_i-\mu_j)(\mathbf{x}_i-\mu_j)^{\mathsf{T}},\\
\frac{\partial \log|\mathbf{\Sigma}|}{\partial \mathbf{\Sigma^{-1}}} &= -\frac{\partial \log|\mathbf{\Sigma^{-1}}|}{\partial \mathbf{\Sigma^{-1}}} = -\mathbf{\Sigma},\\
\frac{\partial (\mathbf{x}_i-\mu_j)^{\mathsf{T}}\mathbf{\Sigma^{-1}}\boldsymbol{\Lambda}_j\mathbf{b}_{ij}^{(t)}}{\partial \mathbf{\Sigma^{-1}}} &= (\mathbf{x}_i-\mu_j)\left(\mathbf{b}_{ij}^{(t)}\right)^{\mathsf{T}}\Lambda_j^{\mathsf{T}},\\
\frac{\partial \mathbf{tr}(\boldsymbol{\Lambda}_j^{\mathsf{T}}\mathbf{\Sigma^{-1}}\boldsymbol{\Lambda}_j\mathbf{C}_{ij}^{(t)})}{\partial \mathbf{\Sigma^{-1}}} &= \boldsymbol{\Lambda}_j\mathbf{C}_{ij}^{(t)}\boldsymbol{\Lambda}_j^{\mathsf{T}}.
\end{aligned}$$

The solution of $\dfrac{\partial Q(\mathbf{\Sigma})}{\partial \mathbf{\Sigma^{-1}}}=0$ yields the maximiser of $Q(\mathbf{\Sigma})$ which is given by

$$\mathbf{\Sigma}=\frac{1}{n}\mathsf{diag}\left[\sum_{i=1}^{n}\sum_{j=1}^{k}\mathsf{a}_{ij}^{(t)}\left((\mathbf{x}_i-\mu_j)\left((\mathbf{x}_i-\mu_j)^{\mathsf{T}}-2\left[\Lambda_j\mathbf{b}_{ij}^{(t)}\right]^{\mathsf{T}}\right)+\boldsymbol{\Lambda}_j\mathbf{C}_{ij}^{(t)}\boldsymbol{\Lambda}_j^{\mathsf{T}}\right)\right].$$

After simplification, the update $\mathbf{\Sigma}^{(t+1)}$ of $\mathbf{\Sigma}$ is now given by

$$\mathbf{\Sigma}^{(t+1)}=\frac{1}{n}\mathsf{diag}\left[\sum_{i=1}^{n}\sum_{j=1}^{k}\mathsf{a}_{ij}^{(t)}\left(\mathbf{x}_i-\boldsymbol{\mu}_j^{(t+1)}-\boldsymbol{\Lambda}_j^{(t+1)}\mathbf{b}_{ij}^{(t)}\right)\left(\mathbf{x}_i-\boldsymbol{\mu}_j^{(t+1)}\right)^{\mathsf{T}}\right].$$

4.7. *Estimating the regression parameters* $\mathbf{\Phi}$

As a function of $\mathbf{\Phi}$, Q can be written as

$$Q(\mathbf{\Phi})=\sum_{i=1}^{n}\sum_{j=1}^{k}\mathsf{a}_{ij}^{(t)}\boldsymbol{w}_i^{\mathsf{T}}\mathbf{\Phi}^{\mathsf{T}}\mathbf{\Psi}^{-\mathbf{1}}\mathbf{b}_{ij}^{(t)}-\frac{1}{2}\sum_{i=1}^{n}\boldsymbol{w}_i^{\mathsf{T}}\mathbf{\Phi}^{\mathsf{T}}\mathbf{\Psi}^{-\mathbf{1}}\mathbf{\Phi}\boldsymbol{w}_i.$$

The maximiser of $Q(\mathbf{\Phi})$ is given by

$$\mathbf{\Phi}^{(t+1)}=\left[\sum_{i=1}^{n}\sum_{j=1}^{k}\mathsf{a}_{ij}^{(t)}\mathbf{b}_{ij}^{(t)}\boldsymbol{w}_i^{\mathsf{T}}\right]\left[\sum_{i'=1}^{n}\boldsymbol{w}_{i'}\boldsymbol{w}_{i'}^{\mathsf{T}}\right]^{-1}.$$

5. Application to synthetic tasks

Our examples in this paper are all based on synthetic datasets. Since our fixed observed covariates are all assumed to be continuous variables, we generate datasets of covariates from multivariate Gaussians with some chosen mean and variance. Once the two sets of covariates are formed, the generation of $\mathbf{x}$ follows easily. As the derivation of our EM algorithm shows, the estimation equations for $\boldsymbol{\mu}$, $\boldsymbol{\Lambda}$, Σ are very much the same as those obtained for the EM for the generic MFA model. On the other hand, the estimation equation for $\mathbf{\Phi}$ is very straightforward. We shall therefore only concentrate on the estimates of $\boldsymbol{\phi}$, since the estimation is done via a new mechanism that we wish to explain and interpret.

5.1. Example 1

We first consider a relatively simple case where the underlying factor model has intrinsic dimensionality $q = 1$. For this toy problem, we choose $p = 3$, $r = 1$, and $s = 2$. Our true parameters are the following: $\boldsymbol{\phi}^{\top} = (3.2, -1.6)$, $\mathbf{\Phi} = 2.7$ and $\Sigma = \mathsf{diag}(0.01, 0.05, 0.02)$. $\Lambda_1 = (0.95, 0.25, 0.55)^{\top}$, $\Lambda_2 = (0.35, 0.95, 0.15)^{\top}$, $\mu_1 = (-2.0, -3.0, -3.7)^{\top}$ and $\mu_2 = (0.0, 0.0, -1.7)^{\top}$. We use the above parameters to generate $n = 255$ observations from a $k = 2$-component mixture of factor analyzers. We also generate the corresponding covariates for $\boldsymbol{z}$ and $\boldsymbol{y}$. In our artificial dataset, we have $n_1 = 204$ and $n_2 = 51$, which translates into the following mixing proportions: $\pi_1 = 0.80$, $\pi_2 = 0.20$.

The application of our estimation scheme to this task yields good results. It is particularly encouraging to point out that the Newton–Raphson iteration used to update the $\boldsymbol{\phi}$ had quadratic local convergence. In fact, in many cases, fewer than 3 Newton–Raphson iterations are required to produce the update $\boldsymbol{\phi}^{(t+1)}$ at each EM iteration, up to a point where one could think of using a *one-step* Newton–Raphson updating instead of *full* Newton–Raphson described earlier. In all our estimations, we use $\boldsymbol{\phi}^{(0)} = (0.0, 0.0)^{\top}$ as our initial guess. Based on the $B = 500$ bootstrap samples used, it is fair to say that the estimates $\widehat{\boldsymbol{\phi}}$ that we obtained are satisfactorily accurate. In fact, the bootstrap estimated average for $\boldsymbol{\phi}$ is $\bar{\widehat{\boldsymbol{\phi}}} = (3.2776, -1.5806)^{\top}$ and the bootstrap estimate of the standard error in this case is $(0.4056, 0.2602)^{\top}$.

5.2. Example 2

Our second example is also a toy problem, with the only difference that we consider more components and more covariates on the component label than earlier. Here, $\boldsymbol{\phi}_1^{\top} = (-1.30, 2.60, -1.25)$, $\boldsymbol{\phi}_2^{\top} = (2.29, -1.40, -2.40)$ and $\boldsymbol{\phi}_3^{\top} = (-1.10, 2.20, -1.30)$. We use $s = 3$, and $k = 4$. Our mixing proportions in this case are $\pi_1 = 0.32$, $\pi_2 = 0.16$, $\pi_3 = 0.23$ and $\pi_4 = 0.29$, which correspond to $n_1 = 160$, $n_2 = 80$, $n_3 = 115$ and $n_4 = 145$ for our sample of $n = 500$ observations. For simplicity, we use zero-vectors as our initial guesses, namely $\boldsymbol{\phi}_1^{(0)} = (0.00, 0.00, 0.00)^{\top}$, $\boldsymbol{\phi}_2^{(0)} = (0.00, 0.00, 0.00)^{\top}$, and $\boldsymbol{\phi}_3^{(0)} = (0.00, 0.00, 0.00)^{\top}$. Using $B = 500$ bootstrap samples as before, we once again obtain satisfactorily accurate parameter estimates. The following table summarises the results obtained on this toy problem. In the table, μ_b and σ_b represent the bootstrap average and standard errors respectively. Once again, it is fair to say that the method yields satisfactorily accurate estimates.

$\boldsymbol{\phi}_1$	$\mu_b(\widehat{\boldsymbol{\phi}_1})$	$\sigma_b(\widehat{\boldsymbol{\phi}_1})$	$\boldsymbol{\phi}_2$	$\mu_b(\widehat{\boldsymbol{\phi}_2})$	$\sigma_b(\widehat{\boldsymbol{\phi}_2})$	$\boldsymbol{\phi}_3$	$\mu_b(\widehat{\boldsymbol{\phi}_3})$	$\sigma_b(\widehat{\boldsymbol{\phi}_3})$
−1.30	−1.21	0.42	2.29	2.39	0.38	−1.10	−0.41	0.34
2.60	2.44	0.29	−1.40	−1.59	0.38	2.20	1.79	0.26
−1.25	−1.18	0.23	−2.40	−2.36	0.30	−1.30	−1.45	0.23

6. Conclusion and discussion

In this paper, we have studied an extension of the MFA model motivated by the possibility that latent variables could be affected by fixed observed covariates. The EM algorithm for this extended model is found to perform well, despite the need for approximate Newton–Raphson updates. Despite some of the weaknesses of the EM algorithm and the Newton–Raphson iterations, the scheme allows us to obtain reasonably accurate parameter estimates. Last but not least, it is worth mentioning that the Newton–Raphson iteration provides an extra advantage in the form of an estimate of the variance–covariance matrix of the maximum likelihood estimate.

While it is possible to extend the covariate mechanism on $\boldsymbol{z}$ by allowing a different $\mathbf{\Phi}_j$ for each component, it must be noted that such an extension could run into greater identifiability problems, partly because of the invariance to permutations of labels.

The MFA model itself already poses two main identifiability problems, one of which is brought about by the factor model, while the other is caused by the invariance of the mixture density to relabelling. Our approach has so far consisted and will once again consist of restricting the model to allow the determination of a unique set of parameters characterising it. In practice, a unique solution is guaranteed by imposing some constraints on Λ so that the only valid solution is the one that satisfies the constraints. For estimability of parameters, constraints are imposed in such a way that the number of parameters to be estimated is at most equal to the number of items of information provided by the sample. Traditionally, there are two types of constraint that are equivalent:

(1) Constrain Λ to be such that $\Lambda^{\mathsf{T}}\Lambda$ is diagonal. Since, $\Lambda^{\mathsf{T}}\Lambda \in \mathbb{R}^{q\times q}$ is symmetric and diagonal, $q(q-1)$ of its elements are all zeros. This means that $q(q-1)$ elements do not need to be estimated by the parameter estimation procedure. This approach is used when estimation is done via a deterministic optimisation algorithm.

(2) A second approach along the lines of [17,20], consists of preassigning values to some entries of Λ. One such constraint proposed by Lopes and West [17] and used in [8] reduces Λ to a block lower diagonal matrix [6] thereby reducing by $\frac{1}{2}q(q-1)$ the number of parameters to be estimated. This is the form of constraints that we use in the Bayesian sampling framework, since its application is straightforward.

Both the above approaches provide an upper bound on the number of factors that can be included in a model. In fact, to guarantee a unique solution under our constraints, all we need is to determine q such that

$$p(q+1) - \tfrac{1}{2}\,q(q-1) \leqslant \tfrac{1}{2}\,p(p+1)$$

which means

$$(p+q) \leqslant (p-q)^2. \tag{30}$$

Note: It must be said that there are situations where solutions satisfying constraint (30) might not provide an adequate fit for the data. In fact, given a data set, a fundamental question without an obvious answer is whether there exists a matrix of factor loadings Λ such that

[6] We assume Λ to be full rank, so we constrain its “diagonal” elements to be nonzero.

the model described by the FA equation adequately fits the data. An exploration of this issue and many other related topics of FA can be found in such references as [3,6,14,15,20] amongst others. Besides the inherent lack of identifiability of the generic MFA model we have to contend here with new aspects of identifiability. As remarked by Titterington [24], it is difficult to give general rules for model identification, so that this difficult issue is always tackled according to the task at hand. Let us consider an unconstrained underlying local FA model, and a $q \times q$ orthogonal transformation Γ such that $\Gamma^{\mathsf{T}}\Gamma = \Gamma\Gamma^{\mathsf{T}} = \mathbf{I}_q$. Given our set $\boldsymbol{\theta} = \{\boldsymbol{\phi}, \mathbf{\Lambda}, \boldsymbol{\mu}, \Sigma, \Phi\}$ of parameters, we apply the following transformations: $\tilde{\mathbf{\Phi}} = \Gamma^{\mathsf{T}}\mathbf{\Phi}$ and $\tilde{\Lambda}_j = \Lambda_j\Gamma$. It is easy to see that both the mean and the covariance matrix in (6) remain unchanged if we substitute Λ_j and $\mathbf{\Phi}$ by $\tilde{\Lambda}_j$ and $\tilde{\mathbf{\Phi}}$, respectively. The parameter set $\tilde{\boldsymbol{\theta}} = \{\boldsymbol{\phi}, \tilde{\mathbf{\Lambda}}, \boldsymbol{\mu}, \Sigma, \tilde{\Phi}\}$ is therefore equivalent to $\boldsymbol{\theta}$, and we conclude that the model as defined is not identifiable. However, if we constrain each local factor analyzer, $\tilde{\boldsymbol{\theta}} = \{\boldsymbol{\phi}, \tilde{\mathbf{\Lambda}}, \boldsymbol{\mu}, \Sigma, \tilde{\Phi}\}$ will define an entirely new model, since transformations will lead to a violation of our restrictions on the structure with parameters not satisfying our constraints. The identifiability of our extended model is therefore achieved by the constraints imposed on the local factor analyzers.

We have so far tested our inference and estimation algorithm only on artificial tasks, but we would like to use it on real life applications. In our future investigations, we plan to address identifiability by implementing a constrained version of the EM algorithm.

A natural alternative to the EM algorithm that we have just studied is the Bayesian treatment of the model. In our analysis of the MFA model, we found that the model allowed the use of conjugate priors, and we used Bayesian sampling on the complete-data posterior to perform estimation and inference. If we consider our set of parameters $\boldsymbol{\theta}$ and the form of the likelihood for the extended MFA model, it is easy to see that we can still use the same priors for $\boldsymbol{\mu}$, $\mathbf{\Lambda}$ and Σ. As far as the two newcomers $\boldsymbol{\phi}$ and $\mathbf{\Phi}$ are concerned, a Gaussian prior on the columns or rows of $\mathbf{\Phi}$ should lead to a full conditional posterior that is also Gaussian. The only parameter that could demand extra concentration of effort in this case is $\boldsymbol{\phi}$. In fact, a good candidate prior for each $\boldsymbol{\phi}_j$ is a Gaussian prior. Let us consider deriving the corresponding full conditional posterior

$$\boldsymbol{p}(\boldsymbol{\phi}_j \mid \cdots) \propto \left[\prod_{i=1}^{n} \left[\boldsymbol{p}(\mathbf{x}_i | \boldsymbol{y}_i, \mathbf{z}_i)\mathbf{Pr}(\boldsymbol{Y}_i = j | \boldsymbol{u}_i)\right]^{y_{ij}}\right] \boldsymbol{p}(\boldsymbol{\phi}_j). \tag{31}$$

In (31), $\boldsymbol{p}(\boldsymbol{\phi}_j)$ is Gaussian, and $\boldsymbol{p}(\mathbf{x}_i | \boldsymbol{y}_i, \mathbf{z}_i)$ is also Gaussian, but the logistic distribution function $\mathbf{Pr}(\boldsymbol{Y}_i = j | \boldsymbol{u}_i)$ is nonGaussian, so that the derivation of $\boldsymbol{p}(\boldsymbol{\phi}_j \mid \cdots)$ is not straightforward. One of the classical solutions to this problem is the use of approximations, namely the Laplace approximation. This Laplace approximation consists of approximating the logistic function by a Gaussian, which then allows the derivation of an approximate Gaussian full conditional posterior $\boldsymbol{p}(\boldsymbol{\phi}_j \mid \cdots)$. This Bayesian treatment will be studied in depth in our future work.

Acknowledgments

The author wishes to thank his Ph.D. supervisor Professor D. M. Titterington for suggesting the problem and for providing helpful guidance and constructive comments.

References

[1] M. Aitkin, G.T. Wilson, Mixture models, outliers, and the em algorithm, Technometrics 22 (1980) 325–331.

[2] G. Arminger, P. Stein, J. Wittenberg, Mixtures of conditional mean and covariance structure models, Psychometrika 65 (1999) 475–494.

[3] D.J. Bartholomew, Latent Variable Models and Factor Analysis, Griffin's Statistical Monographs and Courses, Charles Griffin, London, 1987.

[4] P. McCullagh, J.A. Nelder, Generalized Linear Models, second ed., Monographs on Statistics and Applied Probability, Chapman & Hall, London, 1989.

[5] C. Dolan, H. Van der Maas, Fitting multivariate normal finite mixtures subject to structural equation modelling, Psychometrika 63 (1998) 227–253.

[6] B.S. Everitt, An Introduction to Latent Variable Models, first ed., Monographs on Statistics and Applied Probability, Chapman & Hall, London, 1984.

[7] L. Fahrmeir, G. Tutz, Multivariate Statistical Modelling Based on Generalized Linear Models, Springer Series in Statistics, Springer, Berlin, 1994.

[8] E. Fokoué, Contribution to the analysis of latent structures, Ph.D. Thesis, Department of Statistics, University Glasgow, G12 8QW, UK, 2001.

[9] E. Fokoué, D.M. Titterington, Mixtures of factor analysers: Bayesian estimation and inference by stochastic simulation, Mach. Learning 50 (2003) 73–94.

[10] Z. Ghahramani, M. Beal, Variational inference for Bayesian mixture of factor analysers, in: S.A. Solla, T.K. Leen, K.R. Muller (Eds.), Advances in Neural Information Processing Systems, vol. 12, MIT Press, Cambridge, MA, 2000.

[11] Z. Ghahramani, G.E. Hinton, The EM algorithm for mixtures of factor analyzers, Technical Report CRG-TR-96-1, Department of Computer Science, University of Toronto, Toronto, Canada, M5S 1A4, 1997.

[12] M. Hurn, A. Justel, C. Robert, Estimating mixtures of regressions, Technical report, University of Bath, Department of Mathematical Sciences, 2000.

[13] R.C. Jansen, Maximum likelihood in a generalized linear finite mixture model by using the em algorithm, Biometrics 49 (1993) 227–231.

[14] W. Krzanowski, F. Marriott, Multivariate Analysis, first ed., Kendall's Library of Statistics, vol. 1, Edward Arnold, Paris, 1994.

[15] W. Krzanowski, F. Marriott, Multivariate Analysis, first ed., Kendall's Library of Statistics, vol. 2, Arnold, Paris, 1995.

[16] S. Lee, J. Shi, Bayesian analysis of structural equation model with fixed covariates, Technical report Department of Statistics, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong, 1999.

[17] H.F. Lopes, M. West, Model uncertainty in factor analysis, Technical report ISDS, Institute of Statistics and Decision Sciences, Duke University, 1999.

[18] G. McLachlan, D. Peel, Finite Mixture Models, Wiley Series in Probability and Mathematical Statistics, Wiley, New York, 2000.

[19] B. Muthén, K. Shedden, Finite mixture modelling with mixture Outcomes using the EM algorithm, Biometrics 55 (1999) 463–469.

[20] S.J. Press, Applied Multivariate Analysis, first ed., Holt, Rinehart and Winston, New York, 1972.

[21] M.D. Sammel, L.M. Ryan, J.M. Legler, Latent variable models for mixed discrete and continuous outcomes, J. Roy. Statist. Soc. Ser. B 59 (1997) 667–678.

[22] M.E. Tipping, C.M. Bishop, Mixtures of probabilistic principal component analysers, Neural Comput. 11 (1999) 443–482.

[23] M. Tipping, C. Bishop, Probabilistic principal component analysers, J. Roy. Statist. Soc. Ser. B 61 (1999) 611–622.

[24] D.M. Titterington, A.F.M. Smith, U.E. Makov, Statistical Analysis of Finite Mixture Distributions, Wiley Series in Probability and Mathematical Statistics, Wiley, New York, 1985.

[25] T.J. Thompson, P.J. Smith, J.P. Boyle, Finite mixture models with concomitant information: assessing diagnostic criteria for diabetes, J. Roy. Statist. Soc. Ser. C (Applied Statistics) 47 (1998) 393–404.

[26] N. Ueda, R. Nakano, Z. Ghahramani, E. Hinton, SMEM Algorithm for mixture models, Neural Comput. 12 (2000) 2019–2128.

[27] Y.F. Yung, Finite mixtures in confirmatory factor analysis models, Psychometrika 62 (1997) 297–330.