Electronic Journal of Applied Statistical Analysis
EJASA (2012), Electron. J. App. Stat. Anal., Vol. 5, Issue 3, 346 – 352
e-ISSN 2070-5948, DOI 10.1285/i20705948v5n3p346

BINOMIAL FACTOR ANALYSIS WITH THE MCEM ALGORITHM

Marco Minozzo[*], Clarissa Ferrari

Department of Economics, University of Verona, Italy

Received 17 July 2012; Accepted 20 October 2012
Available online 16 November 2012

***Abstract**: Since its introduction, the classical linear factor model has been central in many fields of application, notably in psychology and sociology, and, assuming continuous and normally distributed observed variables, its likelihood analysis has typically been tackled with the use of the EM algorithm. For the case in which the observed variables are not Gaussian, extensions of this model have been proposed. Here, we present a hierarchical factor model for binomial data for which likelihood inference is carried out through a Monte Carlo EM algorithm. In particular, we discuss some implementations of the estimation procedure with the aim to improve its computational performances. The binomial factor model and the Monte Carlo EM estimation procedure are illustrated on a data set coming from a psychological study on the evaluation of the professional self-efficacy of social workers.*

***Keywords**: Factor analysis, generalized linear mixed models, latent factors, maximum likelihood, Monte Carlo EM.*

1. Introduction

The use of latent variable models is widespread in all fields of applications for the analysis of complex data sets and for the investigation of theoretical hypotheses. In this respect, the classical linear factor model constitutes in many instances the reference tool for the discovery of unobserved variables responsible for the correlation in the observed data, and the expectation-maximization (EM) algorithm provides the basic procedure for the likelihood estimation of its parameters. In social sciences, where the observed (manifest) variables are often non-Gaussian (categorical, for instance), many extensions of the classical factor model have been developed. The work in [1] constitutes one of the main references for the analysis of discrete and categorical multivariate data using latent variable models. More recently, [8] discusses a latent trait model

[*] E-mail: marco.minozzo@univr.it

for mixed outcomes in the class of the generalized linear models, and proposes the use of the Monte Carlo EM (MCEM) algorithm for the inferential analysis of a model with binomial and Gaussian manifest variables. Still adopting the MCEM algorithm, [6] proposes an extension of the classical factor model, with more than one common latent variable, for manifest variables having a distribution in the exponential family. In contrast, for ordered manifest variables and for manifest variables in the exponential family, quite different estimation procedures have been adopted in [5] and in [10] where a marginal likelihood EM approach and a simulated EM algorithm have been applied. Another approach for the analysis of mixed (continuous and discrete) outcomes has been proposed in [9] where polytomous variables are defined via unknown threshold parameters applied to some underlying continuous variables. In a Bayesian inferential context, [2] proposes a mixture of generalized linear models to model the joint distribution of a set of underlying variables related, through distinct link function, to the observed outcomes. Notwithstanding the several proposals appeared in the literature, the computational burden required for the inferential analysis of non-Gaussian factor models is still a challenge. Adopting a likelihood based inferential approach, the stochastic versions of the EM algorithm, such as the MCEM algorithm, seem to work well, but their computational efficiency decrease rapidly as the number of manifest and latent variables increases. In this work, we consider a hierarchical factor model for binomial data and discuss some issues related to the implementation of the MCEM algorithm. The paper is organized as follows. In Section 2 we introduce our binomial factor model, whereas in Section 3 we discuss some aspects related to the implementation of the MCEM algorithm. Then in Section 4 we apply our estimation procedure to a psychological data set, and in Section 5 we end with some conclusions.

2. A hierarchical binomial factor model

Let us present here a hierarchical factor model for multivariate binomial observations. Let y_{ik}, i=1,…,m, k=1,…,K, be a set of real observations (each taking values in {0,1,2,…,n}) relative to K subjects and to m variables (items). For these observations we assume that, for any given i and k, conditionally on the latent random variables Z_{ik}, the random variables Y_{ik} have conditional distributions $f(y;M_{ik})$, that is, $Y_{ik}|Z_{ik} \sim f(y;M_{ik})$, specified by the values of the conditional expectations M_{ik}=E[$Y_{ik}|Z_{ik}$], and that $h(M_{ik})= \beta_i+Z_{ik}$, for some parameters β_i and some known link function $h(\cdot)$. Specifically, we assume that the data are conditionally distributed as Binomial, that is, that, for any i and k,

$$f(y;M_{ik}) = \binom{n}{y}\pi_{ik}^{y}(1-\pi_{ik})^{n-y}, \qquad y=0,1,2,\ldots,n,$$

so that $M_{ik}=n\cdot\pi_{ik}$, and that the linear predictor β_i+Z_{ik} is related to the conditional mean M_{ik} through a logit link function, that is, $\text{logit}(\pi_{ik})=\ln(\pi_{ik}/(1-\pi_{ik}))=\beta_i+Z_{ik}$, and so $M_{ik}=n\cdot e^{\beta i+Zik}/(1+e^{\beta i+Zik})$. For the latent part of the model we adopt the following linear factor structure. For any i and k, let $Z_{ik}=\Sigma^{P}_{p=1}\, a_{ip}\, F_{pk}+\xi_{ik}$, where a_{ip} are $m\times P$ coefficients and the latent variables F_{pk} and ξ_{ik} represent common factors (indexed by p=1,…,P) and unique factors (indexed by i=1,…,m), respectively. We further assume that F_{pk} (responsible for the correlation between the Z_{ik}) are independently and identically distributed (i.i.d.) as standardized Gaussians, and ξ_{ik} are i.i.d. as Gaussians with mean zero and variance equal to ψ_i (Figure 1 shows the conditional independence structure of the model). Let us note that, as the classical factor model, our model is not identifiable. Indeed, for any orthogonal $P \times P$ matrix $\mathbf{G}$ with elements g_{qp} $(q,p=1,\ldots,P)$, such that $\mathbf{G}\mathbf{G}^T=\mathbf{G}^T\mathbf{G}=\mathbf{I}$, we have that the variables $F^{(G)}_{qk}=\Sigma^P_{p=1}\, g_{qp}\, F_{pk}$, $q=1,\ldots,P$, are still Gaussians with mean zero and variance one, so that they are not distinguishable from the common factors F_{pk}, $p=1,\ldots,P$. However, this is the only indeterminacy in the model. As far as the latent variables Z_{ik} are concerned, defining $\varsigma_i^2=\mathrm{Var}[Z_{ik}]$, for $i=1,\ldots,m$, and $\varsigma_{ij}=\mathrm{Cov}[Z_{ik},Z_{jk}]$, for $i\neq j$, we have that $\varsigma_i^2=\Sigma^P_{p=1}\, a_{ip}^2+\psi_i$, and $\varsigma_{ij}=\Sigma^P_{p=1}\, a_{ip}a_{jp}$, for $i\neq j$. This correlation structure is responsible for the correlation between the observable Y_{ik}. Moreover, as in classical linear factor analysis, the coefficients a_{ip} can be interpreted in terms of covariances between the latent Z_{ik} and the common factors F_{pk}. In fact, $\mathrm{Cov}[Z_{ik},F_{pk}]=a_{ip}$, whereas the correlation between Z_{ik} and F_{pk} is given by $\mathrm{Corr}[Z_{ik},F_{pk}]=a_{ip}/\varsigma_i^2$. Defining the standardized loadings $\lambda_{ip}=a_{ip}/\varsigma_i^2$, we can write, for $i\neq j$, $\mathrm{Corr}[Z_{ik},Z_{jk}] = \varsigma_{ij}/(\varsigma_i\, \varsigma_j) = (1/(\varsigma_i\, \varsigma_j))\, \Sigma^P_{p=1}\, a_{ip}\, a_{jp} = \Sigma^P_{p=1}\, \lambda_{ip}\, \lambda_{jp}$, and, by noting that $\mathrm{Corr}[Z_{ik},Z_{ik}] = 1 = \Sigma^P_{p=1}\, (a_{ip}/\varsigma_i)^2 + (\psi_i/\varsigma_i^2) = \Sigma^P_{p=1}\, \lambda_{ip}^2 + (\psi_i/\varsigma_i^2)$, we derive that $0 \leq \Sigma^P_{p=1}\, \lambda_{ip}^2 \leq 1$. Though we are here concerned with Binomial observations, the present model can be adapted to deal with Poisson, Gamma or other kind of manifest variables.

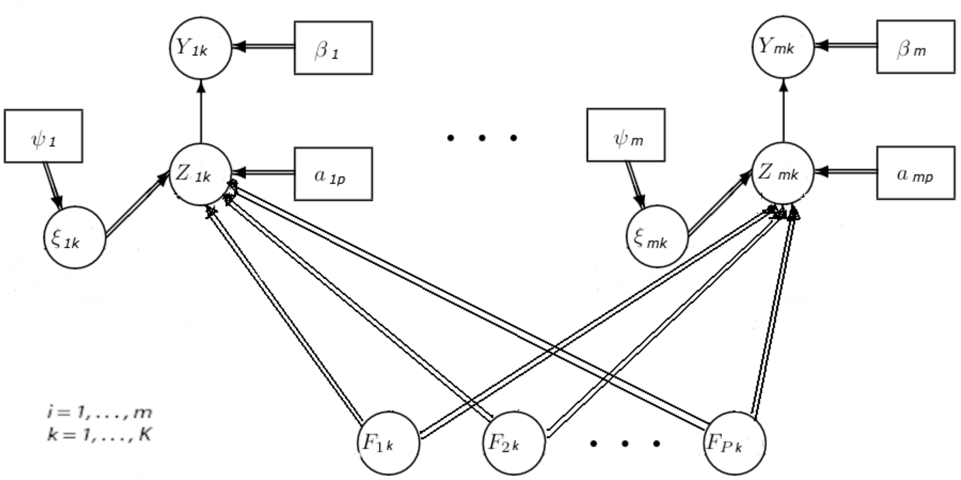

Figure 1. Directed acyclic graph representing the conditional independence structure of the model.

3. Likelihood inference through the MCEM algorithm

The model depends on the parameter vector $\vartheta=(\boldsymbol{\beta},\boldsymbol{\psi},\mathbf{A})$, where $\boldsymbol{\beta}=(\beta_1,\ldots,\beta_m)^T$, $\boldsymbol{\psi}=(\psi_1,\ldots,\psi_m)^T$, and $\mathbf{A}=(\mathbf{a}_1,\ldots,\mathbf{a}_m)^T$, with $\mathbf{a}_i=(a_{i1},\ldots,a_{ip})$, for $i=1,\ldots,m$. For the implementation of the MCEM algorithm we need to work with the complete likelihood (accounting also for the unobserved variables) which can be factorized as:

$$L_c(\vartheta) = f(\mathbf{y},\mathbf{F},\xi) = f(\mathbf{y} \mid \mathbf{F},\xi)\cdot f(\mathbf{F})\cdot f(\xi) = \left[\prod\nolimits_{i=1}^m f(\mathbf{y}_i \mid \mathbf{F},\xi_i)\right]\cdot\left[\prod\nolimits_{p=1}^P f(\mathbf{F}_p)\right]\cdot\left[\prod\nolimits_{i=1}^m f(\xi_i)\right],$$

where $\mathbf{y}=(\mathbf{y}_1,\ldots,\mathbf{y}_m)^T$ with $\mathbf{y}_i=(y_{i1},\ldots,y_{iK})$ for $i=1,\ldots,m$, $\mathbf{F}=(\mathbf{F}_1,\ldots,\mathbf{F}_P)^T$ with $\mathbf{F}_p=(F_{p1},\ldots,F_{pK})$ for $p=1,\ldots,P$, and $\boldsymbol{\xi}=(\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_m)^T$ with $\boldsymbol{\xi}_i=(\xi_{i1},\ldots,\xi_{iK})$ for $i=1,\ldots,m$, and

$$f(\mathbf{F}_p) = \prod_{k=1}^{K} \frac{1}{\sqrt{2\pi}} e^{-F_{pk}^2/2}, \qquad\qquad f(\boldsymbol{\xi}_i) = \prod_{k=1}^{K} \frac{1}{\sqrt{2\pi\psi_i}} e^{-\xi_{ik}^2/(2\psi_i)},$$

$$f(\mathbf{y}_i \mid \mathbf{F}, \boldsymbol{\xi}_i) = \prod_{k=1}^{K} f(y_{ik} \mid F_{1k}, \ldots, F_{Pk}, \xi_{ik}) = \prod_{k=1}^{K} \binom{n}{y_{ik}} \left(\frac{M_{ik}}{n} \right)^{y_{ik}} \left(1 - \frac{M_{ik}}{n} \right)^{n-y_{ik}}$$

$$\propto \prod_{k=1}^{K} \exp\left(y_{ik} \left(\beta_i + \sum_{p=1}^{P} a_{ip} F_{pk} + \xi_{ik} \right) \right) \left(1 + \exp\left(\beta_i + \sum_{p=1}^{P} a_{ip} F_{pk} + \xi_{ik} \right) \right)^{-n}.$$

Taking the logarithm of the complete likelihood we get the complete log-likelihood:

$$l_c(\vartheta) = \sum_{i=1}^{m} l_i(\beta_i, a_{i1}, \ldots, a_{iP}) + \sum_{p=1}^{P} \ln f(\mathbf{F}_p) + \sum_{i=1}^{m} \ln f(\boldsymbol{\xi}_i; \psi_i),$$

where

$$l_i(\beta_i, a_{i1}, \ldots, a_{iP}) \propto \sum_{k=1}^{K} \left[y_{ik} \left(\beta_i + \sum_{p=1}^{P} a_{ip} F_{pk} + \xi_{ik} \right) - n \cdot \ln\left(1 + \exp\left(\beta_i + \sum_{p=1}^{P} a_{ip} F_{pk} + \xi_{ik} \right) \right) \right],$$

$$\ln f(\mathbf{F}_p) = -\frac{K}{2} \ln(2\pi) - \frac{1}{2} \mathbf{F}_p^T \mathbf{F}_p, \qquad p=1,\ldots,P,$$

$$\ln f(\xi_i; \psi_i) = -\frac{K}{2} \ln(2\pi\psi_i) - \frac{1}{2\psi_i} \boldsymbol{\xi}_i^T \boldsymbol{\xi}_i, \qquad i=1,\ldots,m.$$

This complete log-likelihood can be maximized numerically with respect to $\boldsymbol{\vartheta}$. In particular, the maximization with respect to ψ_i, i=1,…,m, (that is, the maximization of $\ln f(\boldsymbol{\xi}_i;\psi_i)$) leads to m closed form formulas, whereas the maximization with respect to β_i and a_{ip}, i=1,…,m and p=1,…,P, can be carried out numerically by maximizing separately the functions $l_i(\beta_i,a_{i1},\ldots,a_{iP})$, i=1,…,m. Let us note that this numerical maximization can be carried out by using standard maximization techniques since each function $l_i(\beta_i,a_{i1},\ldots,a_{iP})$ admits just one point of stationarity, which is a local and global maximum. Indeed, it is easy to see that the Hessian matrices of the functions $l_i(\beta_i,a_{i1},\ldots,a_{iP})$, i=1,…,m, are negative definite (almost surely) for every value of the parameters β_i, $a_{i1},\ldots,a_{iP}$, that is, that the functions $l_i(\beta_i,a_{i1},\ldots,a_{iP})$ are concave with a unique maximum, and that the complete log-likelihood $l_c(\boldsymbol{\vartheta})$ is concave with a unique local and global maximum. On the other hand, in the case in which the loadings $a_{i1},\ldots,a_{iP}$ are known, it is possible to see that the complete likelihood belongs to the curved exponential family and that it is maximized by two continuous functions. In this situations, we can invoke the results in [3] to state that the (stable) MCEM algorithm converges almost surely to some values in the set of stationary points of the likelihood.

4. An application to psychological data

Our binomial factor model and the Monte Carlo EM estimation algorithm are illustrated on a data set coming from a psychological study regarding the evaluation of the professional self-efficacy of social workers. Self-efficacy is defined as a person's belief in her own ability to succeed in specific situations. In professional contexts, a high level of self-efficacy is strictly associated to good professional performances. Self-efficacy can be measured indirectly through ad-hoc questionnaires. Our data is relative to a questionnaire with 11 items, each of which can be answered on a Likert scale made up of 7 ordered categories: the higher the category, the greater the level of self-efficacy. In particular, here we are interested in the analysis of three main aspects of self-efficacy: 'procedural', 'affective' and 'communicative', representing, respectively, the self-efficacy in applying the standard procedures regarding specific situations, the self-efficacy in managing emotions, and the self-efficacy in communicating problems. To this end, we use our model to carry out a factor analysis assuming P=3 latent common factors and m=11 manifest variables (items). Also, since the answer to the 11 items can be given on a Likert scale with 7 categories, we assume for the observed data a Binomial distribution with n=6. The psychological study involved about 800 social workers and for our analysis we have selected 100 questionnaires from the subset (of roughly 770 questionnaires) of those with no missing data. The MCEM estimation algorithm was implemented using the R and OpenBUGS software. Figure 2 shows the run of the MCEM algorithm, whereas Table 1 reports the corresponding estimates.

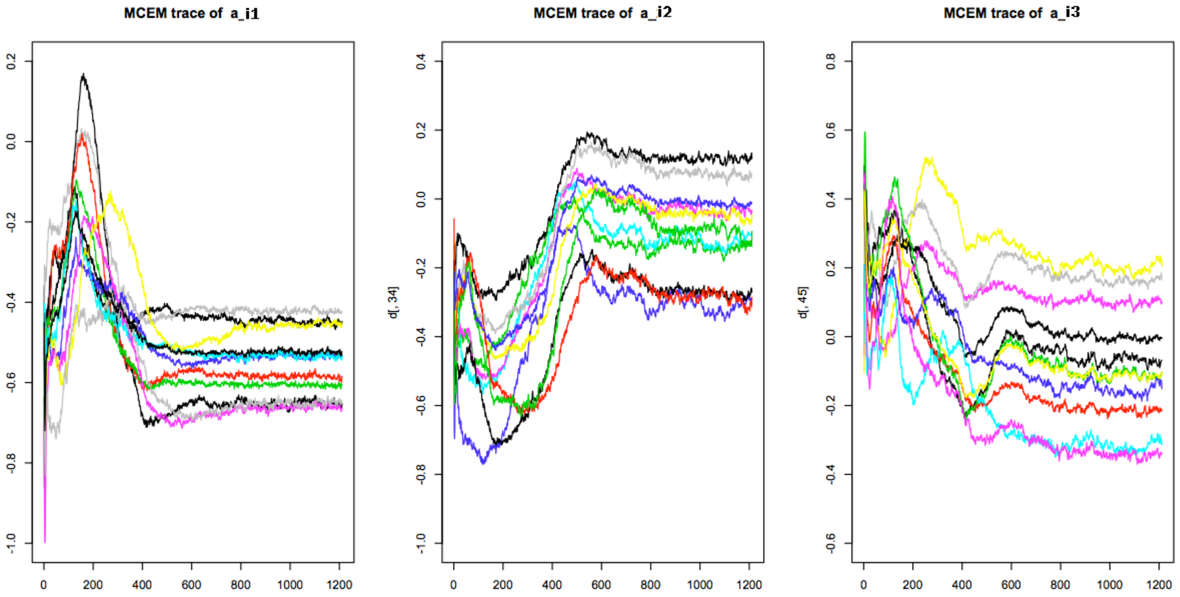

Figure 2. MCEM traces (1200 iterations each involving up to 200 MCMC samples) of the loadings a_{i1}, a_{i2} and a_{i3}, for i=1,…,m, of the Binomial factor model.

As in standard factor analysis, the three common factors extracted can be interpreted looking at the values of the standardized loadings. In particular, the first common factor seems more related to items 1, 5, 6, 7 and 11 and could be named 'procedural' self-efficacy; the second to items 2, 3 and 4; and the third common factor to items 8, 9 and 10, and might represent 'communicative' self-efficacy.

Table 1. Estimated standardized loadings for a model with *m*=11, *P*=3 and *n*=6. For the MCEM algorithm we considered 200 iterations, each involving 200 MCMC samples. Reported estimates have been obtained by averaging over the last 30 iterations and using a varimax rotation.

Item	λ_{i1}	λ_{i2}	λ_{i3}
Item 1 (‘welcome’)	-0.865	-0.393	-0.302
Item 2 (‘managing of impotence’)	-0.370	-0.810	-0.325
Item 3 (‘loading on someone’)	-0.256	-0.864	-0.335
Item 4 (‘containing anxiety’)	-0.469	-0.864	-0.254
Item 5 (‘recognizing limits’)	-0.658	-0.643	-0.390
Item 6 (‘compliance with commitments’)	-0.666	-0.372	-0.644
Item 7 (‘abstaining rating’)	-0.803	-0.417	-0.414
Item 8 (‘needing of support’)	-0.399	-0.565	-0.660
Item 9 (‘seeking of support’)	-0.299	-0.329	-0.895
Item 10 (‘professional involvement’)	-0.620	-0.305	-0.723
Item 11 (‘redefine goals’)	-0.681	-0.466	-0.495

5. Conclusions

The analysis of the above data set with our hierarchical factor model for non-Gaussian data using the MCEM algorithm for the estimation of the parameters gave quite convincing results. Other work is needed to investigate further the convergence and the sampling properties of the MCEM algorithm and to diminish its computational cost. On this last point, it might prove useful to follow some of the suggestions appeared in [4] and [7] which allowed, in other contexts, for a considerable improvement of the computational performances of the estimation procedures.

Acknowledgement

This work is partially based on a presentation given at the SCo2011 conference (Padova, 19-21 September 2011). We gratefully acknowledge funding from the Italian Ministry of Education, University and Research (MIUR) through PRIN 2007 project 20075XA9CJ_005, and PRIN 2008 project 2008MRFM2H.

References

[1]. Bartholomew, D.J. (1980). Factor analysis for categorical data. *Journal of the Royal Statistical Society Series B*, 42, 293–321.

[2]. Dunson, D.B. (2000). Bayesian latent variable models for clustered mixed outcomes. *Journal of the Royal Statistical Society Series B*, 62, 355–366.

[3]. Fort, G., Moulines, E. (2003). Convergence of the Monte Carlo expectation maximization for curved exponential families. *The Annals of Statistics*, 31, 1220-1259.

[4]. Lavine, R.A., Casella, G. (2001). Implementations of the Monte Carlo EM algorithm. *Journal of Computational and Graphical Statistics*, 10, 422–439.

[5]. Moustaki, I. (2000). A latent variable model for ordinal variables. *Applied Psychological Measurement*, 24, 211–223.

[6]. Moustaki, I., Knott, M. (2000). Generalized latent trait models. *Psychometrika*, 65, 391–411.

[7]. Robert, C.P., Casella, G. (2010). *Monte Carlo Statistical Methods*. New York: Springer.

[8]. Sammel, M.D., Ryan, L.M., Legler, J.M. (1997). Latent variable models for mixed discrete and continuous outcomes. *Journal of the Royal Statistical Society Series B*, 59, 667–678.

[9]. Shi, J.Q., Lee, S.Y. (2000). Latent variable models with mixed continuous and polytomous data. *Journal of the Royal Statistical Society Series B*, 62, 77–87.

[10]. Wedel, M., Kamakura, W.A. (2001). Factor analysis with (mixed) observed and latent variables in the exponential family. *Psychometrika*, 66, 515–530.