Estimation Methods for Multivariate Tobit Confirmatory Factor Analysis

D. R. Costa[a], **V. H. Lachos**[a1], **J. L. Bazan**[b] **and C. L. N. Azevedo**[a]

[a] Department of Statistics, Campinas State University, Brazil

[b] Department of Applied Mathematics and Statistics, University of Sao Paulo, Brazil

Abstract

We propose two methods for estimating multivariate Tobit Confirmatory Factor Analysis (TCFA) with covariates, one from Bayesian and another from likelihood based perspectives. TCFA is particularly useful in analysis of multivariate data with censored information. In contrast with previous developments that utilize Monte Carlo simulations for maximum likelihood estimation, an exact EM-type algorithm is proposed, which uses closed form expressions at the E-step that rely on the mean and variance of a truncated multinormal distribution and can be computed using available software. Through simulation studies, we compare the performance of the proposed algorithm when the censored pattern is ignored for different levels of censoring. Our results suggest that this algorithm has excellent performance, since it recovered the true parameters of the TCFA model much better than did the traditional CFA model. In addition, by considering a hierarchical formulation of the models, we also explore the estimation of the parameters via MCMC techniques by using proper priors. A Bayesian case deletion influence diagnostic based on the q-divergence measure and model selection criteria is also developed and applied to analyze a real dataset from an education assessment. In addition, a simulation study is conducted to compare the performance of the proposed method with the traditional CFA model.

Keywords: Multivariate Tobit model, Factor analysis, Latent variable, Bayesian analysis, Censored data.

1 Introduction

Confirmatory factor analysis (CFA) is a type of structural equation modeling (SEM) and has received much attention in recent years (Ullman, 2006). CFA is one of the

[1] *Address for correspondence*: Víctor Hugo Lachos Dávila, Departamento de Estatística, IMECC, Universidade Estadual de Campinas, CEP 13083-859, Campinas, São Paulo, Brazil. E-mail: hlachos@ime.unicamp.br.

most powerful and flexible tools to reduce dimensionality, to describe variability and to model dependency structures in multivariate analysis. Unlike exploratory factor analysis (EFA), CFA provides a more explicit framework for confirming a prior structure of the model (Jreskog, 1969). This technique has been widely used in psychology (Pilati and Laros, 2007), econometrics (Geweke and Singleton, 1981), education testing (Kember and Leung, 1998) and other areas.

CFA models were initially developed for continuous and normally distributed variables. However, some situations require analysis of non-negative data with a large proportion of zeros. By ignoring the information about theses zeros, the analysis will produce a biased representation of the data, since the censoring mechanism that generates the zeros may contain information about the factor structure (Kamakura and Wedel, 2001). Moreover, the results (estimates) obtained by analyzing these kind of data considering standard factor models are biased, since the normality assumption for the response is no longer reasonable (Muthén, 1989). Therefore, Tobit models have been receiving considerable attention lately.

The Tobit model developed by Tobin (1958) has the advantage of providing an explicit link between the data-generating mechanism of both zero and non-zero data by offering a variety of specifications of latent variables and censoring mechanisms and restricting the distribution of the non-censored data to have positive support. Because of its flexibility in modeling this kind of mixed data, the Tobit model has recently attracted much attention in the statistical literature. It was first introduced in the context of factor analysis models by Muthén (1989). Later, Waller and Muthn (1992), Huang (1999), Kamakura and Wedel (2001), Zhou and Liu (2009) developed various procedures to analyze multivariate data through Tobit confirmatory factor analysis (TCFA). Different from these authors, in this article we propose a frequentist estimation methods that does not need Monte Carlo simulation and we also propose a Bayesian framework. Here, the traditional CFA model is a special case (when the information about the censoring threshold is ignored).

The main motivation of the proposed approaches is to present alternative methods of analysis, as well as to provide some additional tools, including influence diagnostic analysis in the TCFA context. The first approach is an exact EM algorithm, where both the E and M steps are performed straightforwardly. In contrast to previous works that implement a type of EM algorithm with techniques of Monte Carlo (see, for example, Huang, 1999; Zhou and Liu, 2009), here we derive closed form expressions at the E step that reduce to computing the first two moments of a truncated multinormal distribution and can be computed using available software, like R (R Core Team, 2012). From this perspective, it is easy to evaluate the likelihood function numerically and, therefore to use it for monitoring convergence and for model selection, such as the Akaike information criterion (AIC), Bayesian information criterion (BIC) and the consistent Akaike information criterion (CAIc).

The second approach is a full Bayesian procedure developed through a MCMC algorithm. Recent developments in Markov chain Monte Carlo (MCMC) methods allow for easy and straightforward implementation of the Bayesian paradigm through conventional software like `OpenBugs`, since suitable hierarchical structures of the model are available. The Bayesian approach allows extreme flexibility in fitting realistic models to datasets of varying complexity (Dunson, 2001), makes use of all information available in the study, accommodates full parameter uncertainty through appropriate prior choices strengthened with proper sensitivity investigations, provides direct probability statements about a parameter through credible intervals, and does not depend on asymptotic results. Additionally, we propose measures of influence diagnostics under the Bayesian paradigm.

The remainder of this paper is organized as follows: in Section 2, the multivariate Tobit latent variable model in confirmatory factor analysis is defined. Section 3 describes the maximum likelihood approach and develops of the exact EM-type algorithm. In Section 4, the Bayesian analysis for this model is defined. The Bayesian case influence diagnostics are presented in Section 5. Both methods are illustrated with the analysis of simulation studies in Section 6 and one example of a real dataset related to educational assessment is presented in Section 7. Finally, Section 8 contains our concluding remarks.

2 The model

The Gaussian confirmatory factor analysis (CFA) model with covariates is specified as follows:

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\Lambda}\mathbf{z}_i + \boldsymbol{\epsilon}_i, \tag{1}$$

where $\mathbf{z}_i \overset{iid}{\sim} N_q(\mathbf{0}, \boldsymbol{\Omega})$ is independent of $\boldsymbol{\epsilon}_i \overset{ind.}{\sim} N_p(\mathbf{0}, \boldsymbol{\Psi})$, $\quad i = 1, \ldots, n$; the subscript i is the subject index; $\mathbf{y}_i = (y_{i1}, \ldots, y_{ip})^\top$ is a $p \times 1$ vector of observed continuous responses for subject i; $\mathbf{X}_i$ is the $p \times k$ design matrix corresponding to the fixed effects represented by:

$$\mathbf{X}_i = \begin{pmatrix} \mathbf{x}_{i1}^\top & & & \\ & \mathbf{x}_{i2}^\top & & \\ & & \cdots & \\ & & & \mathbf{x}_{ip}^\top \end{pmatrix},$$

with $\mathbf{x}_{ij}$ a vector of dimension $k \times 1$; $\boldsymbol{\beta}$, of dimension $k \times 1$; $\boldsymbol{\Lambda}$ is a $p \times q$ loading matrix of factor coefficients; $\mathbf{z}_i$ is a $q \times 1$ $(q < p)$ vector of latent factors ; $\boldsymbol{\epsilon}_i$ of dimension $(p \times 1)$ is the vector of random errors, $\boldsymbol{\Omega}$ is the covariance/correlation matrix of the factors, usually defined beforehand (as a confirmatory analysis), and $\boldsymbol{\Psi}$ is a diagonal covariance matrix of the unobserved errors.

In the present formulation, we consider the case where the response Y_{ij} is not fully observed for all i, j. Following Vaida and Liu (2009) and Matos et al. (2013), let the observed data for the i-th subject be $(\mathbf{V}_i, \mathbf{C}_i)$, where $\mathbf{V}_i$ represents the vector of uncensored reading (i.e., when the observation y_{ij} is not censored, V_{ij} is equal to this observation) or censoring level (when the observation y_{ij} is censored, V_{ij} is a constant), and $\mathbf{C}_i$ is the vector of censoring indicators with components C_{ij}. Then, we have

$$\begin{aligned} y_{ij} &\leq V_{ij} \quad \text{if} \quad C_{ij} = 1, \\ y_{ij} &= V_{ij} \quad \text{if} \quad C_{ij} = 0. \end{aligned} \tag{2}$$

By combining (1) and (2), the multivariate Tobit confirmatory factor analysis (TCFA) model is formulated. Then, the observed value y_{ij} is less than or equal to the censoring level if it is a (left) censored case; otherwise the observed value is not censored. The structure presented in (2) is a generalization of the one discussed in Zhou and Liu (2009), since in our case the censoring level in V_{ij} can assume any value (that is, zero is a special case). The extensions to right or arbitrary censoring are immediate. In general, FA models are not identified, since if we consider $\boldsymbol{\Lambda}^* = \boldsymbol{\Lambda}\boldsymbol{\Gamma}^{-1}$ and $\mathbf{z}_i^* = \boldsymbol{\Gamma}\mathbf{z}_i$, for any nonsingular matrix $\boldsymbol{\Gamma}$, in equation (1), we will obtain the same model. To solve this identification problem, Zhou and Liu (2009), for example, fixed some elements of $\boldsymbol{\Lambda}$ and/or $\boldsymbol{\Omega}$. This is related to the factor rotation issue and it can be helpful to obtain simpler structures in terms of interpretation.

3 Maximum likelihood estimation

The classic inference for the parameter vector $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \boldsymbol{\Lambda}, \Psi)^\top$ is based on the marginal distribution of $\mathbf{y}_i$, $i = 1, \ldots, n$. When the data are not censored, $\mathbf{y}_i \overset{\text{ind.}}{\sim} N_p(\mathbf{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = \boldsymbol{\Lambda\Omega\Lambda}^\top + \boldsymbol{\Psi}$. However, for censored responses, as in (2), we have that $\mathbf{y}_i \sim TN_p(\mathbf{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma}_i; \mathbb{A})$, where $TN_p(.; \mathbb{A})$ denotes the truncated normal distribution on the interval $\mathbb{A}$, where $\mathbb{A}_i = A_{i1} \times \ldots, \times A_{ini}$, with A_{ij} as the interval $(-\infty, \infty)$, if $C_{ij} = 0$ and $(-\infty, 0]$, if $C_{ij} = 1$.

To compute the likelihood function associated with the TCFA model, the first step is to treat separately the observed and censored components of $\mathbf{y}_i$. Following Vaida and Liu (2009) , let $\mathbf{y}_i^o$ be the p^o vector of observed responses and $\mathbf{y}_i^c$ be the p^c vector of censored observations for subject i with $(p = p^o + p^c)$, such that, $C_{ij} = 0$ for all elements in $\mathbf{y}_i^o$, and 1 for all elements in $\mathbf{y}_i^c$. After reordering, $\mathbf{y}_i$, $\mathbf{V}_i$, $\mathbf{X}_i$, and $\boldsymbol{\Sigma}_i$ can be partitioned as

$$\mathbf{y}_i = vec(\mathbf{y}_i^o, \mathbf{y}_i^c),\ \mathbf{V}_i = vec(\mathbf{V}_i^o, \mathbf{V}_i^c),\ \mathbf{X}_i^\top = (\mathbf{X}_i^o, \mathbf{X}_i^c) \text{ and } \boldsymbol{\Sigma}_i = \begin{pmatrix} \boldsymbol{\Sigma}_i^{oo} & \boldsymbol{\Sigma}_i^{oc} \\ \boldsymbol{\Sigma}_i^{co} & \boldsymbol{\Sigma}_i^{cc} \end{pmatrix},$$

where $vec(.)$ denotes the function which stacks vectors or matrices with the same number of columns. Then we have $\mathbf{y}_i^o \sim N_{p^o}(\mathbf{X}_i^o\boldsymbol{\beta}, \boldsymbol{\Sigma}_i^{oo})$, $\mathbf{y}_i^c|\mathbf{y}_i^o \sim N_{p^c}(u_i, \mathbf{S}_i)$, where $\boldsymbol{u_i} = \mathbf{X}_i^c\boldsymbol{\beta} + \boldsymbol{\Sigma}_i^{co}(\boldsymbol{\Sigma}_i^{oo})^{-1}(\mathbf{y}_i^o - \mathbf{X}_i^o\boldsymbol{\beta})$ and $\mathbf{S}_i = \boldsymbol{\Sigma}_i^{cc} - \boldsymbol{\Sigma}_i^{co}(\boldsymbol{\Sigma}_i^{oo})^{-1}\boldsymbol{\Sigma}_i^{oc}$. Now, let $\Phi_p(\mathbf{u}; \mathbf{a}, \mathbf{A})$ and $\phi_p(\mathbf{u}; \mathbf{a}, \mathbf{A})$ be the cdf and pdf, respectively, of $N_p(\mathbf{a}, \mathbf{A})$ computed at $\mathbf{u}$. From Vaida and Liu (2009) and Matos et al. (2013), the likelihood function for cluster i (using conditional probability arguments) is given by

$$\begin{aligned} L_i(\boldsymbol{\theta}) = f(\mathbf{y}_i|\boldsymbol{\theta}) &= P(\mathbf{V}_i|\mathbf{C}_i, \boldsymbol{\theta}) = P(\mathbf{y}_i^c \leq \mathbf{V}_i^c|\mathbf{y}_i^o = \mathbf{V}_i^o, \boldsymbol{\theta})P(\mathbf{y}_i^o = \mathbf{V}_i^o|\boldsymbol{\theta}), \\ &= P(\mathbf{y}_i^c \leq \mathbf{V}_i^c|\mathbf{y}_i^o, \boldsymbol{\theta})f(\mathbf{y}_i^o|\boldsymbol{\theta}) \\ &= \phi_{p^o}(\mathbf{y}_i^o; \mathbf{X}_i^o\boldsymbol{\beta}, \boldsymbol{\Sigma}_i^{oo})\Phi_{p^c}(\mathbf{V}_i^c; u_i, \mathbf{S}_i) = \alpha_i, \end{aligned} \tag{3}$$

which can be evaluated without much computational burden through the `mvtnorm()` routine available in R (see, for example, Genz et al., 2008; R Core Team, 2012). The log-likelihood function for the observed data is thus given by $\ell(\boldsymbol{\theta}|\mathbf{y}) = \sum_{i=1}^n\{\log \alpha_i\}$. The estimates obtained by maximizing the log-likelihood function $\ell(\boldsymbol{\theta}|\mathbf{y})$ are thus the maximum likelihood estimates (MLEs) for the TCFA model defined in (1) and (2).

From equation 3, model selection criteria can be developed to evaluate and compare the performance of the different models. The most used criteria are $AIC =$

$-2\ell(\boldsymbol{\theta}|\mathbf{y}) + 2T$, $BIC = -2\ell(\boldsymbol{\theta}|\mathbf{y}) + Tln(n)$ and $CAIc = -2\ell(\boldsymbol{\theta}|\mathbf{y}) + T(ln(n) + 1)$, where T denotes the number of parameters in the model.

3.1 The EM algorithm

Since the observed log-likelihood function involves complex expressions, it is very difficult to work directly with it. For linear and nonlinear mixed effects models, an EM-type algorithm was developed by Matos et al. (2013) to perform the ML estimation. In this article, we propose a similar EM algorithm for the TCFA model by considering $\mathbf{y}_i$ and $\mathbf{z}_i$ as missing data to update (M-step) all the parameters involved in the model.

Let $\mathbf{y} = (\mathbf{y}_1^\top, \ldots, \mathbf{y}_n^\top)^\top$, $\mathbf{z} = (\mathbf{z}_1^\top, ..., \mathbf{z}_n^\top)^\top$, $\mathbf{V} = vec(\mathbf{V}_1, \ldots, \mathbf{V}_n)$ and $\mathbf{C} = vec(\mathbf{C}_1, \ldots, \mathbf{C}_n)$ such that we observe $(\mathbf{V}_i, \mathbf{C}_i)$ for the i-th subject. In their estimation procedure, $\mathbf{z}$, $\mathbf{Q}$ and $\mathbf{C}$ are treated as hypothetical missing data, and augmented with the observed data set $\mathbf{y}_c = (\mathbf{C}^\top, \mathbf{V}^\top, \mathbf{y}^\top, \mathbf{z}^\top)^\top$. Hence, the EM-type algorithm is applied to the complete-data log-likelihood function $\ell_c(\boldsymbol{\theta}|\mathbf{y}_c) = \sum_{i=1}^n \ell_i(\boldsymbol{\theta}|\mathbf{y}_c)$, where

$$\begin{aligned}\ell_i(\boldsymbol{\theta}|\mathbf{y}_c) &= cte - \frac{1}{2}\left[\log|\boldsymbol{\Psi}| + (\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \boldsymbol{\Lambda}\mathbf{z}_i)^\top \Psi^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \boldsymbol{\Lambda}\mathbf{z}_i) + \log|\boldsymbol{\Omega}|\right] \\ &+ \frac{1}{2}\left[\mathbf{z}_i^\top \boldsymbol{\Omega}^{-1}\mathbf{z}_i\right], \qquad (4)\end{aligned}$$

and cte is a constant that is independent of the parameter vector $\boldsymbol{\theta}$. Given the current estimate $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}^{(k)}$, the E step calculates the conditional expectation of the complete log-likelihood function, given by

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}) = E(\ell_c(\boldsymbol{\theta}|\mathbf{y}_c)|\mathbf{V}, \mathbf{C}, \boldsymbol{\theta}^{(k)}) = \sum_{i=1}^n Q_i(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}) = \sum_{i=1}^n Q_{1i}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}) + \sum_{i=1}^n Q_{2i}(\boldsymbol{\Omega}|\boldsymbol{\theta}^{(k)}),$$

where

$$Q_{1i}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}) = -\frac{1}{2}\left[\log|\Psi| + \widehat{a}_i^{(k)} - 2(\widehat{\mathbf{y}}_i^{k} - \boldsymbol{\Lambda}\widehat{\mathbf{z}}_i^{(k)})\boldsymbol{\beta}^\top \mathbf{X}_i^\top \Psi^{-1} + \mathbf{X}_i\boldsymbol{\beta}\boldsymbol{\beta}^\top \mathbf{X}_i^\top \Psi^{-1}\right]$$

and

$$Q_{2i}(\boldsymbol{\Omega}|\boldsymbol{\theta}^{(k)}) = -\frac{1}{2}\left[\log|\boldsymbol{\Omega}| + tr\left(\boldsymbol{\Omega}^{-1}\widehat{\mathbf{z}_i\mathbf{z}_i^\top}^{(k)}\right)\right] + cte,$$

with $\widehat{a}_i^{(k)} = tr\left(\widehat{\mathbf{y}_i\mathbf{y}_i^\top}^{(k)}\boldsymbol{\Psi}^{-1(k)} - 2\widehat{\mathbf{y}_i\mathbf{z}_i^\top}^{(k)}\boldsymbol{\Lambda}^{\top(k)}\boldsymbol{\Psi}^{-1(k)} + \boldsymbol{\Lambda}^{(k)}\widehat{\mathbf{z}_i\mathbf{z}_i^\top}^{(k)}\boldsymbol{\Lambda}^{\top(k)}\boldsymbol{\Psi}^{-1}\right)$, $\widehat{\mathbf{z}_i\mathbf{z}_i^\top}^{(k)} = E\left[\mathbf{z}_i\mathbf{z}_i^\top|\mathbf{V}_i, \mathbf{C}_i, \boldsymbol{\theta}^{(k)}\right] = \boldsymbol{\Delta}^{(k)} + \boldsymbol{\Delta}^{(k)}\boldsymbol{\Lambda}^{\top(k)}\boldsymbol{\Psi}^{-1(k)}\widehat{b}_i^{(k)}\Psi^{-1(k)}\boldsymbol{\Lambda}^{\top(k)}\boldsymbol{\Delta}^{(k)}$,

$\widehat{b}_i^{(k)} = \left(\widehat{\mathbf{y}_i\mathbf{y}_i^\top}^{(k)} - \widehat{\mathbf{y}_i^{(k)}}\boldsymbol{\beta}^{\top(k)}\mathbf{X}_i^\top - \mathbf{X}_i\boldsymbol{\beta}\widehat{\mathbf{y}_i^{(k)}} + \mathbf{X}_i\boldsymbol{\beta}^{(k)}\boldsymbol{\beta}^{\top(k)}\mathbf{X}_i^\top\right)\boldsymbol{\Delta}^{(k)} = \left(\boldsymbol{\Omega}^{-1} + \boldsymbol{\Lambda}^{\top(k)}\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}\right)$, $\widehat{\mathbf{z}}_i = E\left[\mathbf{z}_i|\mathbf{V}_i, \mathbf{C}_i, \boldsymbol{\theta}^{(k)}\right] = \boldsymbol{\Delta}^{(k)}\boldsymbol{\Lambda}^{t(k)}\boldsymbol{\Psi}^{-1(k)}\left(\widehat{\mathbf{y}_i^k} - \mathbf{X}_i\boldsymbol{\beta}^{(k)}\right)$, $\widehat{\mathbf{y}_i\mathbf{z}_i^\top} = E\left[\mathbf{y}_i\mathbf{z}_i^\top|\mathbf{V}_i, \mathbf{C}_i, \boldsymbol{\theta}^{(k)}\right] = (\widehat{\mathbf{y}_i\mathbf{y}_i^\top}^{(k)} - \widehat{\mathbf{y}_i^{(k)}}\boldsymbol{\beta}^{\top(k)}\mathbf{X}_i^\top)\boldsymbol{\Psi}^{-1(k)}\boldsymbol{\Lambda}^{(k)}\boldsymbol{\Delta}^{(k)}$.

It is clear that the E-step reduces only to the computation of $\widehat{\mathbf{y}_i\mathbf{y}_i^\top}^{(k)} = E\left[\mathbf{y}_i\mathbf{y}_i^\top|\mathbf{V}_i, \mathbf{C}_i, \boldsymbol{\theta}^{(k)}\right]$ and $\widehat{\mathbf{y}}_i^{(k)} = E\left[\mathbf{y}_i|\mathbf{V}_i, \mathbf{C}_i, \boldsymbol{\theta}^{(k)}\right]$, that is, the mean and second moment of a truncated multinormal distribution. These can be determined in closed form, as functions of multinormal probabilities, using a sequence of simple transformations. For more details on the computation of these moments, see Vaida and Liu (2009) and Matos et al. (2013).

Therefore, given the conditional expectations, available from E step, the conditional maximization (CM) steps are described below:

$$\begin{aligned}
\boldsymbol{\beta}^{(k+1)} &= (\sum_{i=1}^{n}\mathbf{X}_i^\top\mathbf{X}_i)^{-1}\sum_{i=1}^{n}\mathbf{X}_i^\top(\widehat{\mathbf{y}}_i^{(k)} - \boldsymbol{\Lambda}^{(k)}\widehat{\mathbf{z}}_i)\\
\boldsymbol{\Lambda}_j^{(k+1)} &= (\sum_{i=1}^{n}\widehat{\mathbf{z}_i\mathbf{z}_i^\top}^{(k)})^{-1}\sum_{i=1}^{n}\left[(\widehat{\mathbf{y}_i\mathbf{z}_i^\top})_j^\top - \mathbf{z}_i(\mathbf{X}_i\boldsymbol{\beta}^{(k)})_j\right]\\
\boldsymbol{\Psi}^{(k+1)} &= diag\left(\frac{\sum_{i=1}^{n}(\mathbf{A}_i + \mathbf{A}_i^\top)}{2n}\right),
\end{aligned}$$

where $\boldsymbol{\Lambda}_j^\top$ is the j-th row of $\boldsymbol{\Lambda}$, $j = 1, \ldots, p$, and $\mathbf{A}_i = \widehat{\mathbf{y}_i\mathbf{y}_i^\top} - 2\widehat{\mathbf{y}_i^{(k)}}\boldsymbol{\beta}^{\top(k)}\mathbf{X}_i^\top - 2\widehat{\mathbf{y}_i\mathbf{z}_i^\top}\boldsymbol{\Lambda}^\top + 2\boldsymbol{\Lambda}^{(k)}\mathbf{z}_i\boldsymbol{\beta}^{\top(k)}\mathbf{X}_i^\top + \mathbf{X}_i\boldsymbol{\beta}^{(k)}\boldsymbol{\beta}^{\top(k)}\mathbf{X}_i^\top + \boldsymbol{\Lambda}^{(k)}\widehat{\mathbf{z}_i\mathbf{z}_i^\top}^{(k)}\boldsymbol{\Lambda}^{\top(k)}$. This process is iterated until some distance involving two successive evaluations of the actual log-likelihood $\ell(\boldsymbol{\theta}|\mathbf{y})$, like $|\ell(\widehat{\boldsymbol{\theta}}^{(k+1)}) - \ell(\widehat{\boldsymbol{\theta}}^{(k)})|$ or $|\ell(\widehat{\boldsymbol{\theta}}^{(k+1)})/\ell(\widehat{\boldsymbol{\theta}}^{(k)}) - 1|$, is small enough. Following Hughes (1999), Matos et al. (2013) and Vaida and Liu (2009), the variance of the fixed effects can be calculated as

$$Var(\widehat{\boldsymbol{\beta}}) = \left(\sum_{i=1}^{n}\mathbf{X}_i^\top\boldsymbol{\Sigma}_i^{-1}\mathbf{X}_i - \mathbf{X}_i^\top\boldsymbol{\Sigma}_i^{-1}Var(\mathbf{y}_i|\mathbf{V}_i, \mathbf{C}_i)\boldsymbol{\Sigma}_i^{-1}\mathbf{X}_i\right)^{-1}. \tag{5}$$

In the next section we present the Bayesian inference for the TCFA model.

4 Bayesian estimation

In this section, we propose a Bayesian full modeling of the TCFA, the related case deletion Bayesian influence diagnostics based on the q-divergence measure (Peng and Dey, 1995) and Bayesian model selection criteria. First, we specify distributions for the parameters, and then we obtain the full conditional distributions, which make it possible to use Gibbs sampling. Then, we develop diagnostic measures to evaluate influential observations using a Bayesian framework.

4.1 Prior and posterior distributions

By considering the model specification given in (1)-(2), the Tobit confirmatory factor analysis model can be easily represented as follows:

$$\begin{aligned} \mathbf{y}_i|\mathbf{z}_i, \mathbf{C}_i, \mathbf{V}_i, \boldsymbol{\theta} &\sim TN(\mathbf{x}_{ij}^\top \boldsymbol{\beta} + \boldsymbol{\Lambda}_r^\top \mathbf{z}_i, \Psi jj; (-\infty, \mathbf{q}_i)) \\ \mathbf{z}_i|\mathbf{C}_i, \mathbf{V}_i &\sim N_q(\mathbf{0}, \boldsymbol{\Omega}), \end{aligned} \tag{6}$$

where $\boldsymbol{\Lambda} = (\boldsymbol{\Lambda}_1, \ldots, \boldsymbol{\Lambda}_q)_{p\times q}$ and $\boldsymbol{\Lambda}_r$ is the r-th column of $\boldsymbol{\Lambda}$, $r = 1, \ldots, q$, $i = 1, \ldots, n$, $j = 1, \ldots, p$ and $TN(\mu, \sigma^2; a)$ denotes the truncated normal distribution at the point a and $q_{ij} = 0$ if $C_{ij} = 1$ (censored case), and $q_{ij} = \infty$ if $C_{ij} = \mathbf{0}$ (non-censored case).

In order to completely specify the Bayesian model, we need to consider prior distributions for all parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \boldsymbol{\Lambda}, \boldsymbol{\Psi})$, assuming $\boldsymbol{\Omega}$ known to fix the model's identification problem. A popular choice to ensure posterior propriety in linear mixed models is to consider proper (but diffuse) conditionally conjugated priors, see Hobert and Casella (1996). For the specific TCFA model, the prior distributions chosen are:

$$\boldsymbol{\beta} \sim N_p(\boldsymbol{\beta}_{\mathbf{0}}, \boldsymbol{S}_{\boldsymbol{\beta}}), \tag{7}$$

$$\boldsymbol{\Lambda}_r \sim N_p(\boldsymbol{\Lambda}_{\mathbf{0}}, \boldsymbol{S}_{\boldsymbol{\Lambda}}), r = 1, \ldots, q, \tag{8}$$

$$\Psi_{jj} \sim IGamma(k_0/2, v_0/2), \tag{9}$$

where $IGamma(a, b)$ is the inverse gamma distribution with mean $b/(a-1)$, $a > 1$, and density $IGamma(.|a, b)$. Ψ_{jj} is the j-th element of the diagonal of $\boldsymbol{\Psi}$, $j = 1, \ldots, p$. Hence, the kernel of the posteriori distribution is given by:

$$\begin{aligned} \pi(\boldsymbol{\theta}|\mathbf{y}, \mathbf{z}, \mathbf{C}, \mathbf{V}) &\propto \textstyle\prod_{i=1}^n TN(\mathbf{y}_i|\mathbf{z}_i, \mathbf{C}_i, \mathbf{V}_i, \boldsymbol{\theta}) \prod_{r=1}^q N_q(\boldsymbol{\Lambda}_r|\boldsymbol{\Lambda}_{\mathbf{0}}, \boldsymbol{S}_{\boldsymbol{\Lambda}}) \times \\ &\textstyle\prod_{j=1}^p [N_p(\beta_j|\boldsymbol{\beta}_{\mathbf{0}}, \boldsymbol{S}_{\boldsymbol{\beta}}) IGamma(\Psi_{jj}|k_0, v_0)], \end{aligned} \tag{10}$$

From (6), we have that $\mathbf{z}_i|\mathbf{y}_i, \mathbf{C}_i, \mathbf{V}_i, \boldsymbol{\theta} \sim N_q(\widehat{\mathbf{z}}_i, \boldsymbol{\Delta})$, where $\widehat{\mathbf{z}}_i = \boldsymbol{\Delta}\boldsymbol{\Lambda}^\top \boldsymbol{\Psi}^{(-1)}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})$ and $\boldsymbol{\Delta} = (\boldsymbol{\Omega}^{-1} + \boldsymbol{\Lambda}^\top \boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda})^{-1}$. Since the posterior distribution, given by equation (10), is not analytically tractable, MCMC algorithms can be employed to obtain numerical approximation for all marginal posterior distributions. Let

$\boldsymbol{\theta}_1|\mathbf{y}, \mathbf{C}, \mathbf{V}, \boldsymbol{\theta}_{(-\boldsymbol{\theta}_1)}$ be the full conditional density of $\boldsymbol{\theta}_1$ and, with $\boldsymbol{\theta}_1|\mathbf{y}, \mathbf{C}, \mathbf{V}, \boldsymbol{\theta}_{(-\boldsymbol{\theta}_1)} = \boldsymbol{\theta}_1|\mathbf{y}, \boldsymbol{\theta}_{(-\boldsymbol{\theta}_1)}$. Then, we have:

$$\begin{aligned}
\boldsymbol{\beta}|\mathbf{y}, \mathbf{z}, \boldsymbol{\theta}_{(-\boldsymbol{\beta})} &\sim N\left(\mathbf{A}_\beta u_\beta,\ \mathbf{A}_\beta\right), \\
\Psi_{jj}|\mathbf{y}, \mathbf{z}, \boldsymbol{\theta}_{(-\Psi_{jj})} &\sim IGamma(\frac{q_0+n}{2}, \frac{v_0+s}{2}), \\
\boldsymbol{\Lambda}_r|\mathbf{y}, \mathbf{z}, \boldsymbol{\theta}_{(-\boldsymbol{\Lambda}_r)} &\sim N\left(\mathbf{A}_\Lambda \mathbf{u}_\Lambda,\ \mathbf{A}_\Lambda\right),\ r = 1, ..., q
\end{aligned}$$

where $\mathbf{A}_\beta = (\mathbf{S}_\beta^{-1} + \sum_{i=1}^n \mathbf{X}_i^\top \boldsymbol{\Psi}^{-1} \mathbf{X}_i)^{-1}$, $\boldsymbol{u}_{\boldsymbol{\beta}} = (\boldsymbol{\beta}_{\mathbf{0}}^\top \mathbf{S}_{\boldsymbol{\beta}}^{-\mathbf{1}} + \sum_{i=1}^n (\mathbf{y}_i - \mathbf{X}_i^\top \boldsymbol{\beta}_{\mathbf{0}})^\top \boldsymbol{\Psi}^{-\mathbf{1}} \mathbf{X}_i^\top)$, $s = \sum_{i=1}^n (\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta} - \Lambda_r^\top \mathbf{z}_i)$, $\mathbf{A}_\Lambda = (\sum_{i=1}^n \mathbf{z}_i \mathbf{z}_i^\top \boldsymbol{\Psi}^{-1} + \boldsymbol{\Lambda}_0^{-1})^{-1}$, $\boldsymbol{u}_{\boldsymbol{\Lambda}} = \sum_{i=1}^n \mathbf{A}_{ir}^\top \boldsymbol{\Psi}^{-\mathbf{1}} \mathbf{z}_{ir} + \boldsymbol{\Lambda}_{\mathbf{0}}^\top \mathbf{S}_{\boldsymbol{\Lambda}}^{-\mathbf{1}}$, $\mathbf{A}_{ir} = \mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta} - \sum_{l \neq r}^{p_{l=1}} \Lambda_l \mathbf{z}_{il}$, $r = 1, ..., q$.

Therefore, the MCMC algorithm simulates iteratively from the above full conditional distributions. In the next subsection we present some discussion on model comparison criteria. For further details we refer to Lachos et al. (2013).

4.2 Bayesian model comparison criteria

We use the conditional predictive ordinate (CPO) statistic, one of the most widely used model selection/assessment criteria available in the Bayesian toolbox, which is derived from the posterior predictive distribution (see Carlin and Louis, 2008). Let $\mathcal{D}$ be the full dataset and $\mathcal{D}^{(-i)}$ stand for the dataset with the ith observation deleted. We denote the posterior density of θ given $\mathcal{D}^{(-i)}$ by $\pi(\theta|\mathcal{D}^{(-i)})$. For the i-th observation, the CPO_i can be written as $CPO_i = \int_\Theta f(\mathbf{y}_i|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{\mathcal{D}}^{(-i)})d\boldsymbol{\theta} = \left\{\int_\Theta \frac{\pi(\boldsymbol{\theta}|\boldsymbol{\mathcal{D}})}{f(\mathbf{y}_i|\boldsymbol{\theta})} d\boldsymbol{\theta}\right\}^{-1}$. For our proposed model, a closed form of the CPO_i is not available. However, a Monte Carlo estimate of CPO_i can be obtained by using a single MCMC sample from the posterior distribution $\pi(\boldsymbol{\theta}|\mathcal{D})$ using a harmonic-mean approximation, see Dey et al. (1997), that is $\widehat{CPO}_i = \left\{\frac{1}{Q}\sum_{q=1}^{Q} \frac{1}{f(\mathbf{y}_i|\boldsymbol{\theta}_q)}\right\}^{-1}$, where $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_Q$ is a valid MCMC sample of size Q (i.e., disregarding the simulated values before the burn-in and taking spaced observations) from $\pi(\theta|\mathcal{D})$. A summary of the CPO_is is the log pseudo-marginal likelihood (LPML), defined by $LPML = \sum_{i=1}^{n} \log(\widehat{CPO}_i)$. The larger the value of $LPML$, the better the quality of model fit. Other measures, such as the deviance information criterion (DIC) proposed by Spiegelhalter et al. (2002), the expected Akaike information criterion (EAIC) and the expected Bayesian (or Schwarz) information criterion (EBIC) as given in Carlin and Louis (2001), can also be used. These measures are based on the posterior mean of the deviance, which can be approximated by $\overline{D} = \sum_{q=1}^{Q} D(\boldsymbol{\theta}_q)/Q$, where $D(\boldsymbol{\theta}) = -2\sum_{i=1}^{n}\log\left[f(\mathbf{y}_i|\theta)\right]$. The DIC can be estimated using the MCMC output as $\widehat{DIC} = \overline{D} + \widehat{\rho_D}$, where $\rho_D = E\{D(\boldsymbol{\theta})\} - D\{E(\boldsymbol{\theta})\}$ is the effective number of parameters and $D\{E(\boldsymbol{\theta})\}$ is the deviance evaluated at the posterior mean. Similarly, the EAIC and EBIC can be estimated as $\widehat{EAIC} = \overline{D} + 2\#(\vartheta)$ and $\widehat{EBIC} = \overline{D} + \#(\vartheta)\log(n)$, where $\#(\vartheta)$ is the number of parameters in the model. Note that for all these criteria, the evaluation of the likelihood function given in (3) is a key aspect. However, it can be easily computed from our proposed methods, treating separately the observed and censored components of $\mathbf{y}_i$ as presented in Subsection 3.

5 Bayesian case influence diagnostics

Inferences in a factor analysis model with covariates, as in any regression model, can be strongly affected by the inclusion or deletion of a small set of observations. To study the effects of an influential observation on the analysis, pertubation schemes have been developed in the literature (see Cook (1986)). The most commonly used schemes are based on case deletion (Cook and Weisberg, 1982), in which the effects of completely removing cases from the analysis are studied. For our model, we will consider a case-deletion scheme for Bayesian analysis based on the use of perturbation functions.

Perturbation functions were introduced by Kass et al. (1989) and Weiss (1996). From these functions it is possible to evaluate the influence of the assumptions of model M on the posterior distribution $\pi\left(\boldsymbol{\theta}|\mathbf{y}, M\right)$. Suppose that $\pi\left(\boldsymbol{\theta}|\mathbf{y}, M_1\right)$ is the posterior distribution of $\boldsymbol{\theta}$ under model M_1 and $\pi\left(\boldsymbol{\theta}|\mathbf{y}, M_2\right)$ is its posterior distribution under model M_2. Therefore, the perturbation function for case deletion is defined by $p(\boldsymbol{\theta}) = \frac{\pi\left(\boldsymbol{\theta}|\mathbf{y}, M_2\right)}{\pi\left(\boldsymbol{\theta}|\mathbf{y}, M_1\right)}$.

Let us consider a subset I with k elements of the set $\{1, \ldots, n\}$. When the subset I is deleted from the dataset $\mathbf{y}$, we denote the eliminated data as $\mathbf{y}_I$ and $\mathbf{y}_{(-I)}$ the remaining data. Then, the perturbation function for deletion cases is $p(\boldsymbol{\theta}) = \pi\left(\boldsymbol{\theta}|\mathbf{y}_{(-I)}\right)/\pi\left(\boldsymbol{\theta}|\mathbf{y}\right)$. After some straightforward algebraic manipulations, the perturbation function can be defined as

$$p(\boldsymbol{\theta}) = \frac{\left[\prod_{i\in I} f(\mathbf{y}_i|\boldsymbol{\theta})\right]^{-1}}{E_{\boldsymbol{\theta}|\mathbf{y}}\left\{\left[\prod_{i\in I} f(\mathbf{y}_i|\boldsymbol{\theta})\right]^{-1}\right\}}, \tag{11}$$

where $f(\mathbf{y}_i|\boldsymbol{\theta})$ represents the likelihood given in Equation 3.

The perturbation function for the parameters of the TCFA model for the deleted cases can be approximated by using the marginal distribution of $\mathbf{y}$ and MCMC techniques by sampling from the posterior distribution. In fact, when subset $I = \{i\}$ is considered and $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_Q$ is a valid MCMC sample of size Q of $\pi(\boldsymbol{\theta} \mid \mathbf{y})$, the MC approximation of the perturbation function $p(\boldsymbol{\theta})$ is given by

$$\widehat{p(\boldsymbol{\theta})} = \widehat{\mathrm{CPO}_i}\left[\phi_{p^o}(\mathbf{y}_i^o; \mathbf{X}_i^o\boldsymbol{\beta}, \boldsymbol{\Sigma}_i^{oo})\Phi_{p^c}(\mathbf{V}_i^c; u_i, \mathbf{S}_i)\right]^{-1}, \tag{12}$$

Another common approach to quantifying influential observations is using divergence measures between posterior distributions with and without a given subset of the data. The q-divergence measure between two densities π_1 and π_2 for $\boldsymbol{\theta}$ is defined by Csiszár (1967) as

$$d_q(\pi_1, \pi_2) = \int q\left(\frac{\pi_1(\boldsymbol{\theta})}{\pi_2(\boldsymbol{\theta})}\right)\pi_2(\boldsymbol{\theta})d\boldsymbol{\theta}, \tag{13}$$

where q is a convex function such that $q(1) = 0$. Specific divergence measures are obtained by considering particular $q(\cdot)$ functions. For example, the Kullback-Leibler divergence is obtained when $q(z) = -\log(z)$, the J-distance (symmetric version of Kullback-Leibler divergence) is obtained when $q(z) = (z-1)\log(z)$ and the L_1-distance arises by taking $q(z) = |z-1|$ (Lachos et al., 2013).

The q-influence of the data $\mathbf{y}_I$ on the posterior distribution of $\boldsymbol{\theta}$, $d_q(I) = d_q(\pi_1, \pi_2)$, is obtained by considering $\pi_1(\boldsymbol{\theta}) = \pi_1(\boldsymbol{\theta}|\mathbf{y}_{(-I)})$ and $\pi_2(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta}|\mathbf{y})$ in Equation (13), and can be written as

$$d_q(I) = E_{\boldsymbol{\theta}|\mathbf{y}}\left[q(p(\boldsymbol{\theta}))\right], \tag{14}$$

where the expected value is taken with respect to the unperturbed posterior distribution. These influence measures have already been used by Peng and Dey (1995) and Weiss (1996) and more recently by Vidal and Castro (2010).

Note that the influence measure $d_q(I)$ itself does not determine when an observation is influential. It is necessary to define a cutoff point to determine whether a small subset of observations is influential. In this context, we will use the proposal given by Peng and Dey (1995).

Thus, we consider the probability function of a biased coin, which is given by $\pi_1(x \mid p) = p^x(1-p)^{1-x}$, with $x = 0, 1$, while the probability function of an unbiased coin is given by $\pi_2(x \mid p) = 0.5$. From (13), the q-divergence between a biased and an unbiased coin is given by

$$d_q(p) = \frac{q(2p) + q(2(1-p))}{2},$$

where $d_q(p)$ increases as p moves away from 0.5, is symmetric around $p = 0.5$ and achieves its minimum value at $p = 0.5$. Also, if $d_q(0.5) = 0$ then $\pi_1 = \pi_2$. Consequently, if we consider $p \geq 0.85$ (or $p \leq 0.15$) as a strong bias in a coin, then, $d_{L_1}(0.85) = 0.70$ and we can indicate an influential observation when $d_{L_1}(i) \geq 0.70$, $i = 1, \ldots, n$. Similarly, for the Kullback-Leibler divergence we have $d_{KL}(0.85) = 0.33$, and for the J-distance $d_J(0.85) = 0.61$. These cutoff values will be used in this work.

In the next sections, we illustrate the performance of the proposed methods with the analysis of artificial examples and the analysis of a real dataset.

6 Simulated data

In order to examine the performance of the proposed methods, here we report two simulation studies. The first part is devoted to comparing the estimates obtained from both TCFA and CFA models in datasets with different levels of censoring from both frequentist and Bayesian perspectives. The goal of the second part is to show that ML estimates based on the proposed EM algorithm have good asymptotic properties.

We performed simulations from the model defined in (1) and (2), with $p = 5$, $k = 3$ and $q = 2$. The true parameters of $\boldsymbol{\beta}$, $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$ are set at:

$$\begin{aligned}
\boldsymbol{\beta}^\top &= (0.5, 0.8, -0.5) \\
\boldsymbol{\Lambda}^\top &= \begin{pmatrix} -0.6 & -0.6 & -0.6 & 0 & 0 \\ 0 & 0 & 0 & 0.5 & 0.5 \end{pmatrix}, \\
\boldsymbol{\Psi} &= diag(0.2, 0.2, 0.2, 0.2, 0.2)
\end{aligned}$$

As in a confirmatory analysis, the $\boldsymbol{\Omega}$ was fixed at pre-assigned values $0.6\mathbf{I} + 0.4\mathbf{J}$, where $\mathbf{I}$ is a 2×2 identity matrix and $\mathbf{J}$ is a 2×2 matrix of ones. The initial values for both studies were:

$$\begin{aligned}
\boldsymbol{\beta}^{(0)\top} &= (-0.4, -0.4, 0.5) \\
\boldsymbol{\Lambda}^{(0)\top} &= \begin{pmatrix} -0.1 & -0.1 & -0.1 & 0 & 0 \\ 0 & 0 & 0 & 0.1 & 0.1 \end{pmatrix}, \\
\boldsymbol{\Psi}^{(0)} &= diag(0.4, 0.4, 0.4, 0.4, 0.4)
\end{aligned}$$

The matrix of covariates $\mathbf{X}_i$ was generated by the Kronecker product of A and B, where A is a $1 \times p$ matrix of ones and B is a vector of the transposed $\mathbf{x}_{ij}$, $j = 1, 2, 3$. The covariate x_{i1} was equal to 1 for all $i = 1, \ldots, n$ (intercept); x_{i2} was generated from a normal distribution with mean 6 and variance 1 and finally x_{i3} was independently generated from a Bernoulli(0.5) distribution. All programs were implemented in the R software (R Core Team, 2012). In all situations, we compared the consequences on parameter inference when the mechanism of censorship is taken into consideration (TCFA model) or ignored (CFA model).

6.1 Parameter recovery under the TCFA and CFA models

The main focus of this simulation study is to investigate the effect of the level of censoring on the estimation using the proposed EM and Bayesian MCMC methods. We chose three values of censoring proportions (5%, 10% and 50%) and we estimated the parameters of both TCFA and CFA models from a sample size equal to 150. For all scenarios, we simulated 100 datasets (replicas) and calculated the Monte Carlo mean (MC mean) and standard deviations (MC Sd) from these simulated samples.

From a Bayesian perspective, we used the following prior distributions for the parameters: $\beta_l \sim N_1(1, 10^2)$, $\Lambda_{1k} \sim N_3(-0.8, 1), k = 1, 2, 3$, $\Lambda_{1k} \sim N_2(0, 10^{-6}), k = 3, 4$, $\Lambda_{2k} \sim N_3(0, 10^{-6})$, $k = 1, 2, 3$, $\Lambda_{2k} \sim N_2(0.8, 1)$ $k = 3, 4$ and $\Psi_l \sim IGamma(2, 1)$. The use of informative priors was necessary to identify the model. For each sample, we generated two parallel independent MCMC runs of size 20,000, where the first 2,000 iterations (burn-in samples) were discarded, to compute posterior estimates. To eliminate potential problems due to autocorrelation, we considered a lag of size 20. The convergence of the MCMC chains was monitored using trace plots, autocorrelation (ACF) plots and Gelman-Rubin $\hat{R}$ diagnostics. We fit the models using the `R2OpenBUGS` package available in the *R* system.

Table 1 gives some summary statistics of the parameters from all scenarios. We observe that for all levels of censoring patterns and type of analysis (classic or Bayesian), the TCFA model outperforms the CFA one. Since the datasets were generated from a TCFA model, this study shows that TCFA is robust even with a higher percentage of censored cases, providing more accurate estimates than the CFA model. Also, the results show that the proposed methods, under frequentist and Bayesian perspectives, performance well in order to recover the true parameters.

Table 1: Simulated data. Comparison of the estimates from the proposed EM algorithm and Bayesian analysis for different levels of censoring. MC mean and MC Sd are the respective Monte Carlo mean and standard deviations from the 100 replicas. The values in parentheses are the standard errors of the fixed effects using Equation 5. MC mean and MC Sd are the Monte Carlo mean and standard deviations, respectively, from the 100 replicas.

Censoring	Parameters	EM results				Bayesian results			
		TCFA		CFA		TCFA		CFA	
		MC Mean	MC Sd	MC Mean	MC Sd	MC Mean	MC Sd	MC Mean	MC Sd
5%	β_1	0.491 (0.214)	0.189	0.671 (0.205)	0.182	0.475	0.198	0.708	0.187
	β_2	0.802 (0.035)	0.031	0.774 (0.034)	0.030	0.804	0.033	0.767	0.031
	β_3	-0.507 (0.073)	0.063	-0.494 (0.071)	0.060	-0.498	0.061	-0.483	0.059
	λ_{11}	-0.577	0.052	-0.555	0.049	-0.609	0.058	-0.581	0.054
	λ_{21}	-0.577	0.057	-0.556	0.055	-0.608	0.058	-0.584	0.055
	λ_{31}	-0.582	0.055	-0.560	0.051	-0.606	0.058	-0.579	0.055
	λ_{42}	0.463	0.049	0.451	0.045	0.495	0.065	0.479	0.062
	λ_{52}	0.462	0.043	0.449	0.042	0.495	0.064	0.477	0.061
	ψ_{11}	0.199	0.036	0.185	0.033	0.215	0.037	0.200	0.034
	ψ_{22}	0.196	0.033	0.182	0.030	0.212	0.037	0.197	0.033
	ψ_{33}	0.203	0.035	0.188	0.033	0.220	0.037	0.204	0.034
	ψ_{44}	0.215	0.033	0.201	0.031	0.222	0.048	0.209	0.044
	ψ_{55}	0.213	0.027	0.199	0.024	0.217	0.047	0.205	0.043
10%	β_1	0.489 (0.217)	0.193	0.876 (0.200)	0.182	0.472	0.203	0.959	0.181
	β_2	0.802 (0.035)	0.032	0.742 (0.033)	0.03	0.804	0.034	0.727	0.030
	β_3	-0.507 (0.074)	0.062	-0.479 (0.069)	0.058	-0.499	0.061	-0.464	0.057
	λ_{11}	-0.576	0.053	-0.538	0.048	-0.605	0.059	-0.559	0.053
	λ_{21}	-0.577	0.059	-0.540	0.053	-0.610	0.059	-0.564	0.053
	λ_{31}	-0.581	0.054	-0.542	0.049	-0.605	0.060	-0.559	0.053
	λ_{42}	0.463	0.049	0.442	0.044	0.498	0.066	0.465	0.060
	λ_{52}	0.461	0.046	0.438	0.042	0.495	0.066	0.463	0.059
	ψ_{11}	0.200	0.036	0.173	0.031	0.216	0.038	0.189	0.031
	ψ_{22}	0.196	0.033	0.170	0.029	0.212	0.038	0.185	0.031
	ψ_{33}	0.204	0.037	0.176	0.032	0.221	0.039	0.193	0.032
	ψ_{44}	0.216	0.034	0.188	0.029	0.222	0.049	0.197	0.041
	ψ_{55}	0.214	0.026	0.185	0.022	0.220	0.048	0.195	0.040
50%	β_1	0.209 (0.270)	0.31	3.040 (0.157)	0.157	0.382	0.299	3.172	0.139
	β_2	0.843 (0.043)	0.046	0.432 (0.026)	0.024	0.817	0.047	0.401	0.023
	β_3	-0.518 (0.086)	0.08	-0.310 (0.054)	0.042	-0.508	0.075	-0.279	0.044
	λ_{11}	-0.618	0.081	-0.421	0.048	-0.612	0.077	-0.390	0.058
	λ_{21}	-0.625	0.082	-0.424	0.042	-0.619	0.077	-0.393	0.057
	λ_{31}	-0.623	0.074	-0.423	0.042	-0.617	0.078	-0.390	0.059
	λ_{42}	0.498	0.069	0.368	0.038	0.425	0.137	0.291	0.059
	λ_{52}	0.493	0.067	0.362	0.038	0.429	0.134	0.293	0.058
	ψ_{11}	0.204	0.043	0.088	0.019	0.232	0.051	0.106	0.017
	ψ_{22}	0.195	0.043	0.086	0.019	0.226	0.050	0.103	0.016
	ψ_{33}	0.209	0.047	0.091	0.020	0.237	0.052	0.107	0.017
	ψ_{44}	0.220	0.044	0.097	0.018	0.240	0.061	0.115	0.021
	ψ_{55}	0.217	0.039	0.093	0.017	0.234	0.060	0.113	0.020

6.2 Large sample properties of ML estimates

In order to study some large sample proprieties of ML estimates, we fixed the sample size at $n = 30, 50, 150, 350$ and 600. For each sample size, 100 samples from the TCFA model were generated and estimated from the frequentist framework based on the proposed EM algorithm. For this example, the censoring level was fixed at 20%. Comparing the results of CFA with covariates (excluding censorship) and the TCFA for the same dataset, we obtained the estimates presented in Figures 1 to 6. As general rule, we can say that bias and MSE tend to approach zero when the sample size increases, indicating that the ML estimates based on the proposed EM-type algorithm present good large sample properties. We can also highlight that the TCFA model was more stable and yielded more accurate estimation in relation to the analysis that ignores censorship, since the two statistics tended to decrease monotonically with increasing sample size.

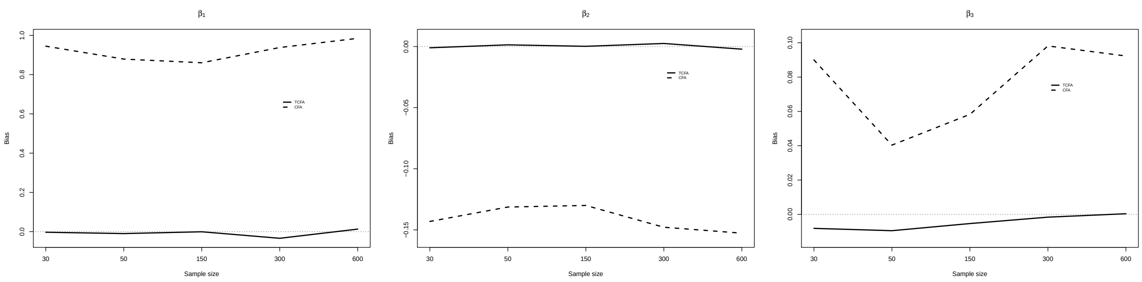

Figure 1: Simulated data. Bias from the parameter $\boldsymbol{\beta}$.

7 Real data analysis

In this section we present an application of the TCFA model in an education assessment. The data refer to the Early Grade Reading Assessment (EGRA), which is a tool used to measure students' reading progress (RTI International, 2009). Specifically, the EGRA assesses how well children in the early grades of primary school are acquiring key reading skills and also determines which areas of instruction need improvement. This test is applied in Latin America and the Caribbean and is administered orally, one student at a time. In about 15 minutes, it examines a student's ability to perform fundamental prereading and reading skills. For our study, we used the EGRA results for Peruvian students in four out of ten subtests/tasks. This base

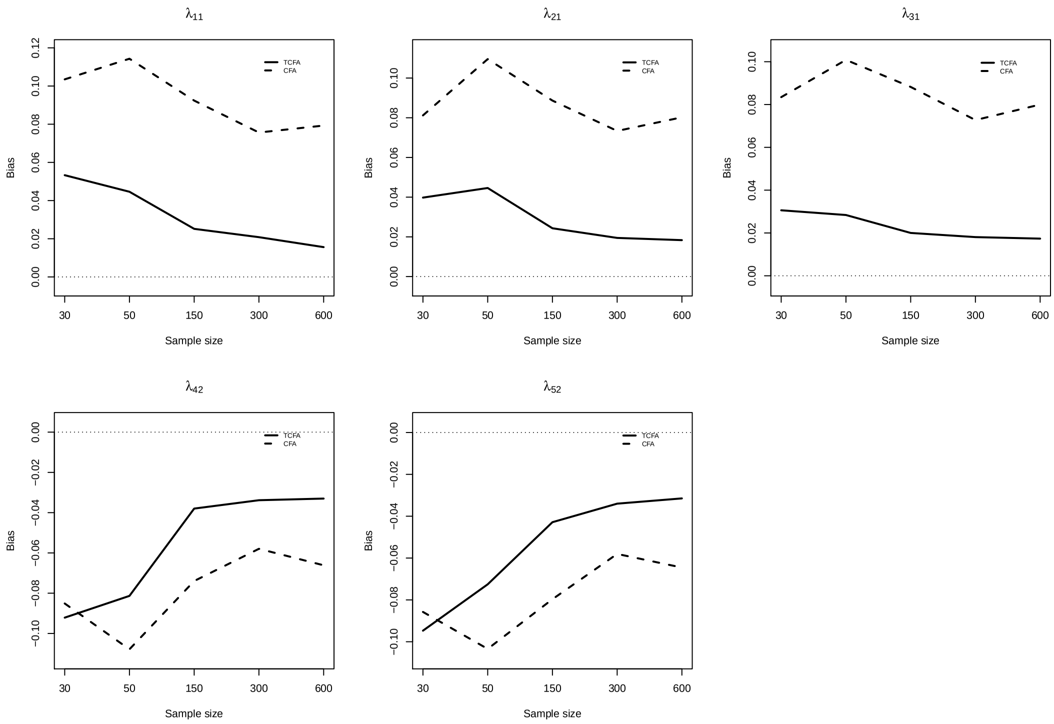

Figure 2: Simulated data. Bias from the parameter $\mathbf{\Lambda}$.

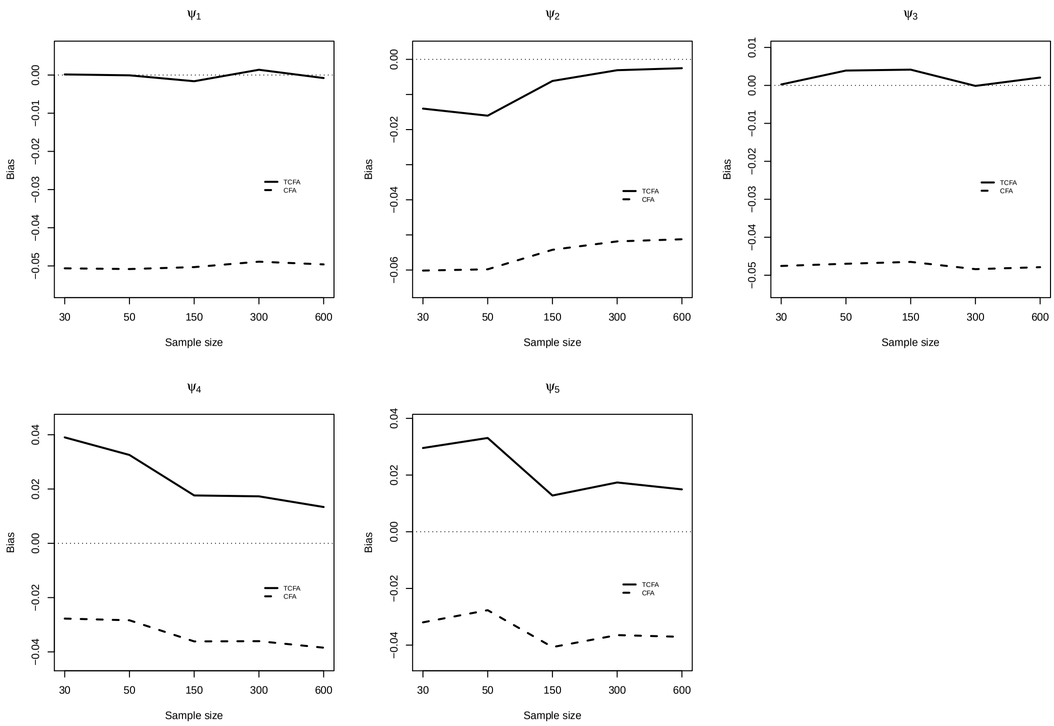

Figure 3: Simulated data. Bias from the parameter $\mathbf{\Psi}$.

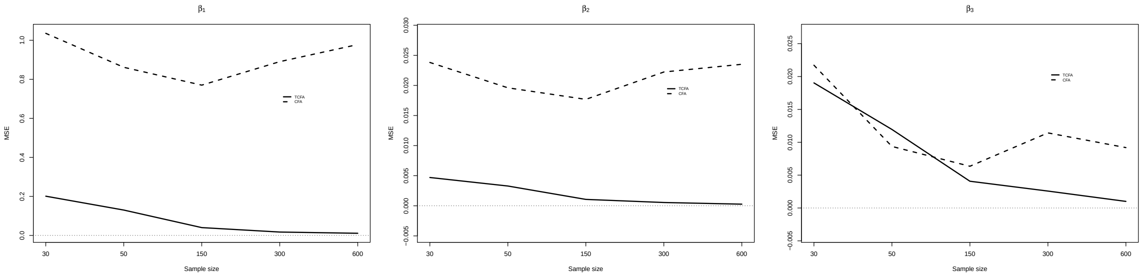

Figure 4: Simulated data. MSE from the parameter $\boldsymbol{\beta}$.

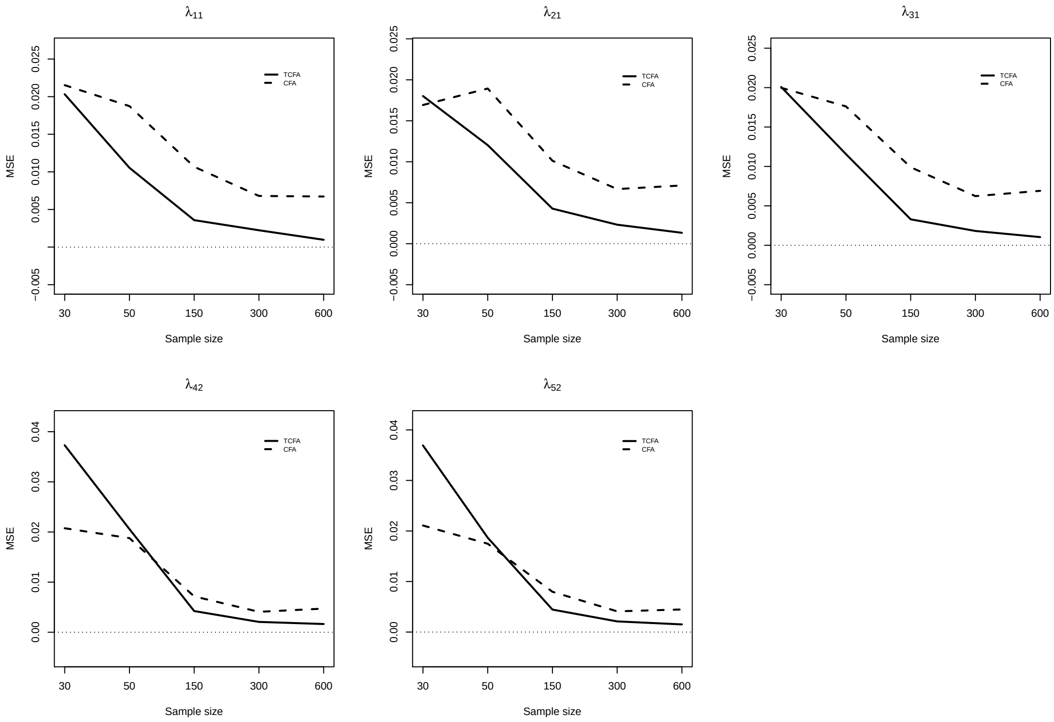

Figure 5: Simulated data. MSE from the parameter $\boldsymbol{\Lambda}$.

refers to the evaluation carried out in 2007. The four analyzed tasks are described in Table 2.

Using this dataset, the main objective is to investigate whether these four subtests can better represent a more general ability: fluency in Spanish. Since the nature of these tasks that take into account a specific time, the scores of the students were transformed in a scale of velocity in the following way: $\text{Velocity}_{ij} = \frac{Y_{ij}}{Time_{ij}}$, where: Y_{ij}=number of letters/words read by student i in task j within 60 seconds or less and $Time_{ij}$= time (in seconds) spent by student i in test j (less than or equal to 60). This transformation indicates that students with high score in fluency in Spanish

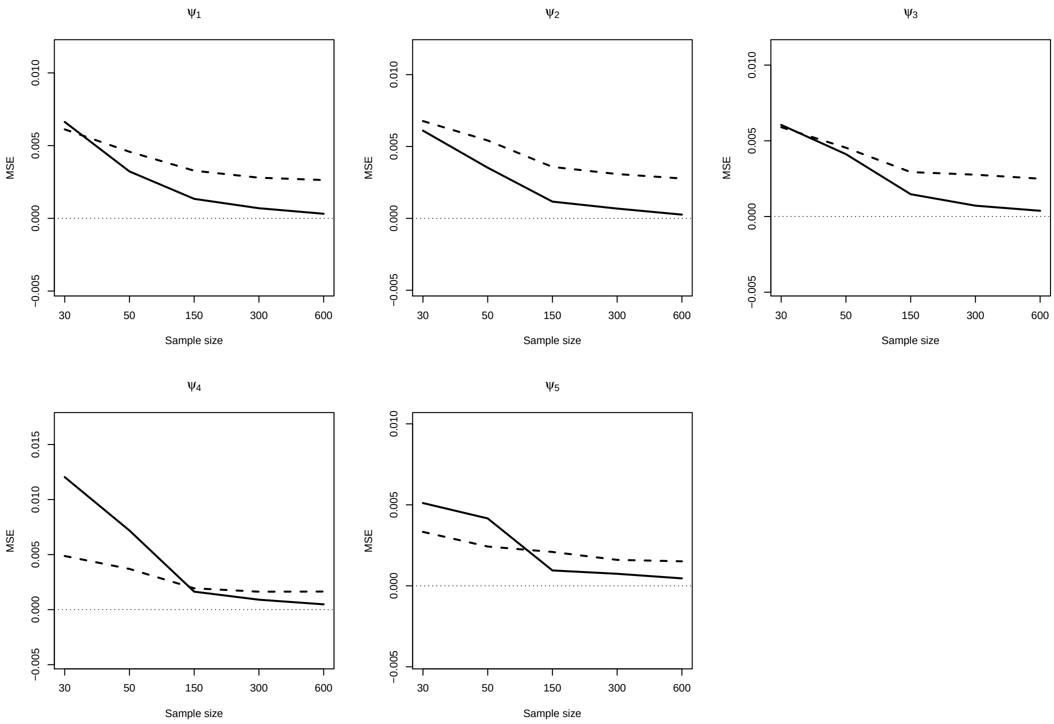

Figure 6: Simulated data. MSE from the parameter $\mathbf{\Psi}$.

will be faster than a student with average or low fluency.

It is important to stress that students with slower velocity measure (inferior to the 10% of the total lower scores) are set equal to zero in that measure (censored outcome). We considered this type of censoring under the assumption that the time of that specific task was not sufficient to better estimate the responses of these students. Using this definition of censored observations, the TCFA model was defined by

$$\mathbf{Velocity}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{\Lambda}\mathbf{z}_i + \boldsymbol{\epsilon}_i, \tag{15}$$

where $\mathbf{Velocity}_i = (Velocity_{i1}, \ldots, Velocity_{i4})^\top$ is a 4×1 vector of the velocity responses for student i on the four tasks, $i = 1, \ldots, 502$; $\mathbf{X}_i$ is the 4×5 design matrix corresponding to the fixed effects defined by β_1 =gender(0=Female, 1=Male); β_2 =grade (0=Second year; 1=Third year); β_3 =residence zone (0=Rural, 1=Urban);β_4 =age; $\mathbf{\Lambda}$ is a (4×1) vector of factor loadings; $\mathbf{z}_i$ is a latent factor associated with the general ability (fluency in Spanish); $\boldsymbol{\epsilon}_i$ of dimension (4×1) is the vector of random errors, and Ω is a scalar (in our case, fixed at 1).

The data apply to 502 students (157 girls and 345 boys; 354 second graders and 148 third graders; 250 from urban zones and 252 from rural ones; 51% seven years

Table 2: EGRA data. Descriptions of the tasks.

Task	Ability measured	Variable
Task 1	Recognizing letters of the alphabet	Y_1=number of correct letters in 1 minute
Task 2	Recognizing simple words	Y_2=number of correct readings of simple words in 1 minute
Task 3	Simple decodification of meaningless words	Y_3=number of correct readings of meaningless words in 1 minute
Task 4	Reading of a passage	Y_4=number of correct readings of simple words in the passage in 1 minute

old or less). As expected for models under factor analysis, we observe moderate to high correlations among response variables from this dataset. Like in the simulation studies, we compared the analysis including and excluding censorship. The ML estimates for $\boldsymbol{\beta}$ (with respective standard error), $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$, fixed $\Omega = 1$, $p = 4$, $q = 1$ and $k = 4$, are summarized in Table 3. The response variables were standardized.

We also estimated the TCFA model through the Bayesian framework for this dataset. Following Lopes and West (2004), we considered the following independent prior distributions in R: $\beta \sim N(0, 10), \Lambda \sim N(0, 1)\mathbf{1}(\lambda_i > 0)$ and $\Psi \sim IGamma(1.1, 0.05)$.

The results of 50,000 iterations in two parallel independent MCMC chains are summarized in Table 3. We discarded first 5,000 iterations (burn-in samples) for computing posterior estimates and, to eliminate potential problems due to auto-correlation, we considered a lag of size 20. The convergence of the MCMC chains was monitored using trace plots, auto-correlation (ACF) plots and Gelman-Rubin $\hat{R}$ diagnostics.

Table 3 shows the difference of the estimates for both analyses (CFA and TCFA models) and from both methods (frequentist and Bayesian). The regression coefficients $\boldsymbol{\beta}$ are used to analyze the effect of covariates x_i on the velocity response. Comparing the frequentist and Bayesian perspectives, in general, the posterior means of parameters are close to the respective ML estimates. From Bayesian point of view, we also added in Table 3 the 95% credibility intervals. Despite the standardization of the response variables, the estimates show that younger third-grade boys from urban zone were faster than other students. We can see that all the variables were significantly different from zero in any scenario.

From an inspection of the results presented in Table 3, we can see that the estimates of the factor loadings and the specific variances (matrix $\boldsymbol{\Psi}$) are relatively

Table 3: EGRA data. Comparison of estimates of the parameters from the proposed EM algorithm and Bayesian analysis and from both TCFA and CFA models.

Parameters	EM results		Bayesian results						
	TCFA	CFA		TCFA			CFA		
	Estimates	Estimates		Mean	Sd	CI(95%)	Mean	Sd	CI(95%)
β_1	0.251 (0.093)	0.295 (0.087)		0.199	0.074	(0.054;0.344)	0.244	0.067	(0.113;0.374)
β_2	0.765 (0.119)	0.736 (0.112)		1.108	0.095	(0.922;1.294)	1.097	0.086	(0.929;1.266)
β_3	0.931 (0.079)	0.938 (0.074)		0.863	0.062	(0.742;0.985)	0.894	0.056	(0.783;1.004)
β_4	-0.123 (0.060)	-0.122 (0.056)		-0.215	0.047	(-0.308;-0.123)	-0.208	0.042	(-0.292;-0.125)
λ_{11}	0.565	0.540		0.580	0.047	(0.489;0.674)	0.544	0.044	(0.459;0.631)
λ_{21}	0.967	0.884		0.985	0.040	(0.909;1.065)	0.879	0.033	(0.816;0.946)
λ_{31}	0.915	0.837		0.925	0.040	(0.850;1.006)	0.828	0.034	(0.763;0.897)
λ_{41}	0.961	0.885		0.973	0.041	(0.896;1.056)	0.876	0.035	(0.810;0.946)
ψ_{11}	0.788	0.721		0.806	0.055	(0.706;0.920)	0.734	0.048	(0.646;0.834)
ψ_{22}	0.161	0.140		0.157	0.019	(0.122;0.196)	0.139	0.016	(0.109;0.171)
ψ_{33}	0.224	0.201		0.228	0.021	(0.188;0.273)	0.203	0.018	(0.169;0.240)
ψ_{44}	0.206	0.179		0.207	0.020	(0.168;0.249)	0.179	0.017	(0.148;0.215)
Model comparison criteria									
Loglik	-2103.952	-2121.841	LPML	-2111.687			-2128.43		
AIC	4231.904	4267.682	DIC	4170.288			4200.523		
BIC	4282.527	4318.305	EAIC	4219.165			4249.831		
CAIc	4294.527	4330.305	EBIC	4286.423			4317.089		

higher for the model that includes censorship. Large values of the factor loadings confirm our initial hypothesis of a general latent factor. This main factor was interpreted as "factor of fluency in Spanish". The variable with less impact on the estimation of this general factor and higher variability ($\Psi_{44} = 0.788$) was the Task 1 (Recognizing letters of the alphabet). Although obtain similar results, we could have different conclusions in terms of number of underlying latent factors from CFA and TCFA models. When we compare the models, all analysed criteria select the TCFA model. These results reinforce the importance of including the information about censorship in the data analysis.

We also analyzed Bayesian case influence diagnostics using the d_q divergence. The results are shown in Figure 7. The first three plots present the measures K-L divergence, J-distance and L_1-distance when the information about the censorship is ignored and the remaining three when the censored cases are taken into consideration.

From Figure 7, the performance of two students ($\#316, \#244$) were indicated as atypical cases in the dataset according to the majority of the influence measures studied. Student #289 was also flagged as influential cases in accordance with

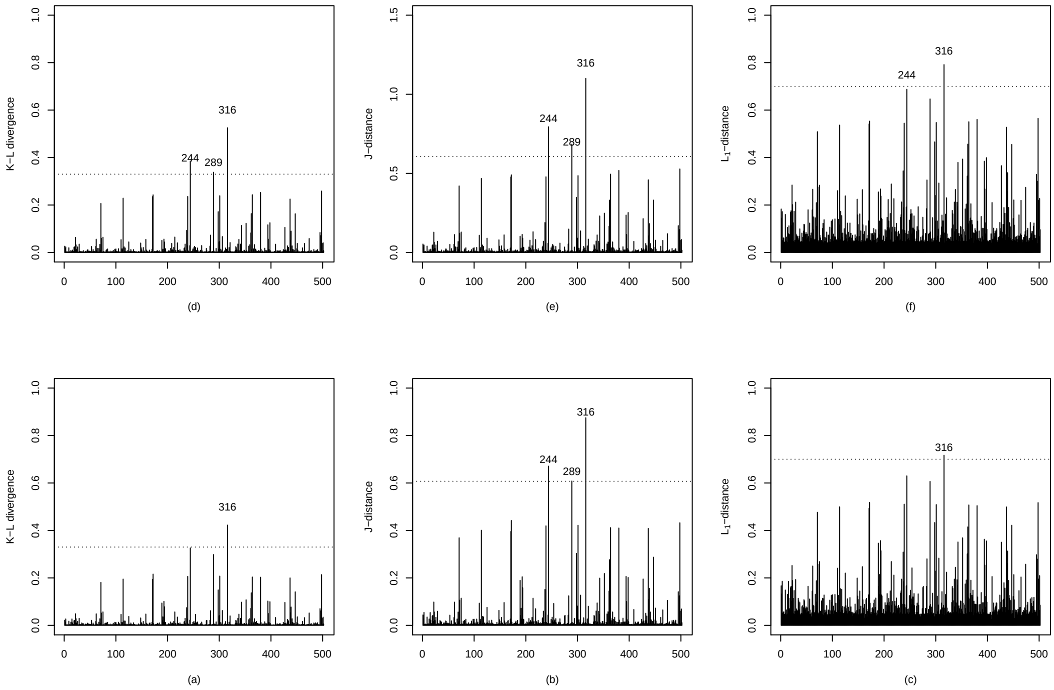

Figure 7: Index plots of d_q measure for the EGRA data. The first three plots (a-c) refer to the CFA model and the last three (d-f) to the TCFA.

$J-$distance. Table 4 shows the characteristics of these students in the dataset.

Table 4: EGRA data. Performance of the four observations indicated as atypical cases according to the influence measures. The first, second and third quantiles of the velocity responses (in second) are also presented.

Observations	Gender	Grade	Zone	Age	Velocity scores (in seconds)			
					Task 1	Task 2	Task 3	Task 4
244	Female	3rd year	Urban	7	2.767	1.724	2.381	1.449
289	Female	2nd year	Urban	8	1.667	3.333	1.367	1.971
316	Female	2nd year	Urban	8	1.500	3.333	2.222	1.818
Quantiles								
25 %					0.700	1.167	0.767	1.190
50 %					0.967	1.533	1.033	1.667
75 %					1.300	1.887	1.358	2.128

The three observations indicated as most influential under the analyzed measures were girls from urban zone. While students #289 and #316 were in the second year of school, the other one was in third year. The youngest age was student #244, 7 years old. With respect to the four tasks in the test, these students had unusual behavior in relation to the other students. Student #244 had good performance on tasks 1 e 3 (in the top 25 % of scores in this task) and medium to weak performance in other tasks. Student #316 had high performance in tasks 2 and 3 and weak velocity measure in tasks 1 e 4. Student #289, on the other hand, had lower performance on Task 3, but his velocity measurements are still configured in the 25% higher scores.

Figure 8 shows the individual profiles of all students in each task in the EGRA data. The three observations that stood out according to the analysis of influence diagnostic measures are highlighted.

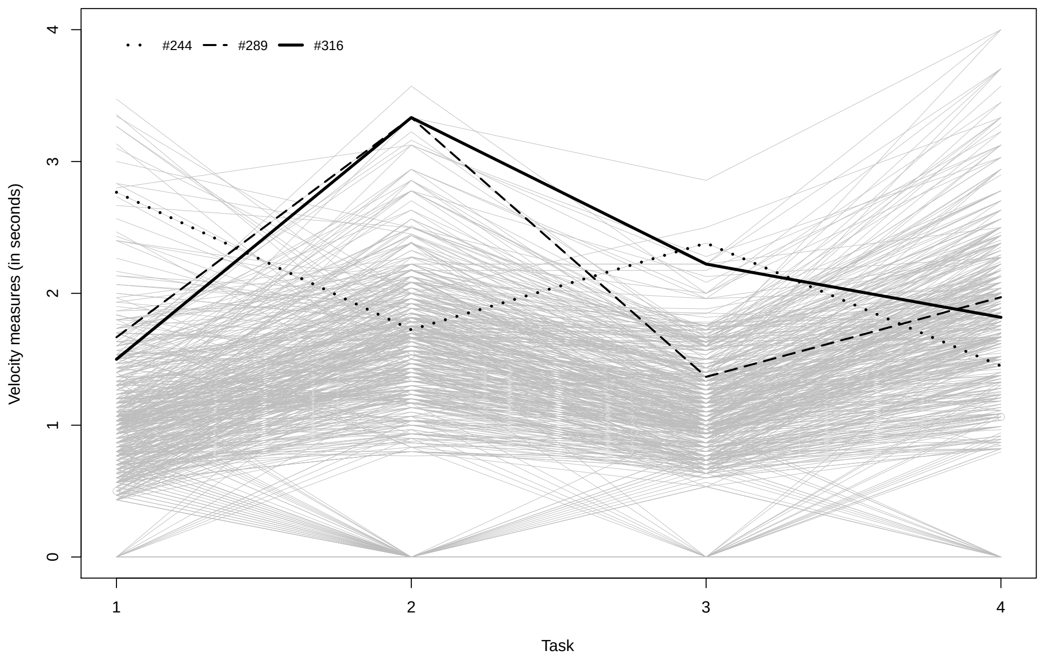

Figure 8: EGRA data. Individual velocity measures for each task. The trajectories for the influential individuals are numbered.

In order to reveal the impact of the two highest influential measure cases (observations #244 and #316) as potentially influential on the estimates of the parameters of the TCFA model, we refitted this model dropping each of these cases. Table 5 presents the relative changes (RC) in percentage of these estimates defined by

$$RC_{\boldsymbol{\theta}} = \left| \frac{\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_{(I)}}{\hat{\theta}} \right| \times 100,$$

where $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\Lambda}}, \hat{\boldsymbol{\Psi}})$ and $\hat{\boldsymbol{\theta}}_{(I)}$ denotes the estimate of parameters after the set I of observations has been removed for both analysis (frequentist and Bayesian).

Table 5: Relative changes (in percentage) in TCFA model for EGRA's data.

RC (in %)	EM algorithm			Bayesian analysis		
	A−{244}	A−{316}	A−{244, 316}	A−{244}	A−{316}	A−{244, 316}
$RC_{\hat{\beta}_1}$	7.190	3.860	11.366	3.695	4.117	8.331
$RC_{\hat{\beta}_2}$	3.644	0.534	4.349	0.577	0.498	0.092
$RC_{\hat{\beta}_3}$	1.275	0.597	1.902	0.263	1.088	0.847
$RC_{\hat{\beta}_4}$	13.719	0.904	13.283	1.804	1.534	0.068
$RC_{\hat{\lambda}_1}$	0.827	0.749	0.109	1.210	0.748	0.421
$RC_{\hat{\lambda}_2}$	0.394	0.414	0.810	0.023	0.541	0.554
$RC_{\hat{\lambda}_3}$	0.691	0.674	0.039	1.253	0.762	0.490
$RC_{\hat{\lambda}_4}$	0.105	0.607	0.492	0.207	0.724	0.914
$RC_{\hat{\Psi}_1}$	0.302	0.438	0.718	0.391	0.527	0.886
$RC_{\hat{\Psi}_2}$	1.793	0.838	1.090	1.988	0.450	1.588
$RC_{\hat{\Psi}_3}$	5.040	1.953	3.236	4.910	1.832	3.105
$RC_{\hat{\Psi}_4}$	2.067	6.351	8.414	2.225	6.570	8.851

From Table 5, we note that the most impact on the estimates, when observations #244 and #316 were dropped, were at: $\hat{\beta}_4$ (age), $\hat{\beta}_1$ (gender), $\hat{\Psi}_3$ and $\hat{\Psi}_4$. The frequentist and Bayesian analysis were coherent, with close estimates. Besides some large RC values, from both perspectives, the parameter significance and sign of the coefficients remained the same when the observation set (#244, #316) was eliminated.

8 Conclusions

In this article we proposed a frequentist and also a Bayesian framework for estimating multivariate Tobit confirmatory factor analysis with covariates. These approaches are good alternatives to better estimate parameters from a multivariate dataset with large percentage of censoring in the data. From the classic approach, we obtained the ML estimates of the model by applying an exact EM algorithm, with its E step made feasible by formulas for the mean and variance of a truncated multinormal distribution. This algorithm circumvents direct evaluation of the intractable observed likelihood function and, as expected, it was numerically stable with increasing likelihood as the number the iterations grew. From a Bayesian standpoint, we proposed a hierarchical formulation of the TCFA for censored responses. These approaches were applied in simulation studies and to a real dataset. We also developed tools for detecting influential observations using q-divergence measures, and quantified their effects on the posterior estimates of model parameters. Under censored data, the results show that the proposed method is efficient and it is encouraging that the use of both methods offer more precise inferences than the traditional CFA, which ignores the information about the censoring threshold. Finally, the proposed algorithms were coded and implemented in R software (R Core Team, 2012) and are available from us upon request.

There are a number of possible extensions of the current work. For instance, it is of interest to generalize the TCFA model by incorporating the multivariate Student-t distribution (see Matos et al., 2013) and to make allowance for missing data. An in-depth investigation of such extensions is beyond the scope of the present paper, but is an interesting topic for further research.

Acknowledgements

This study was supported by FAPESP, CAPES and CNPq, Brazil.

References

Carlin, B., Louis, T., 2008. Bayesian Methods for Data Analysis (Texts in Statistical Science). Chapman and Hall/CRC, New York,.

Carlin, B. P., Louis, T. A., 2001. Bayes and Empirical Bayes Methods for Data Analysis, 2nd Edition. Chapman & Hall/CRC, Boca Raton.

Cook, R. D., 1986. Assessment of local influence. Journal of the Royal Statistical Society, Series B 48, 133–169.

Cook, R. D., Weisberg, S., 1982. Residuals and Influence in Regression. Chapman & Hall/CRC, Boca Raton, FL.

Csiszár, I., 1967. Information-type measures of difference of probability distributions and indirect observations. Studia Scientiarum Mathematicarum Hungarica 2, 299–318.

Dey, D. K., Chen, M. H., Chang, H., 1997. Bayesian approach for the nonlinear random effects models. Biometrics 53, 1239–1252.

Dunson, D., 2001. Commentary: practical advantages of Bayesian analysis of epidemiologic data. American Journal of Epidemiology 153 (12), 1222.

Genz, A., Bretz, F., Hothorn, T., Miwa, T., Mi, X., Leisch, F., Scheipl, F., 2008. mvtnorm: Multivariate Normal and t Distribution. R package version 0.9-2, URL http://CRAN. R-project. org/package= mvtnorm.

Geweke, J. F., Singleton, K. J., 1981. Maximum likelihood" confirmatory" factor analysis of economic time series. International Economic Review 22 (1), 37–54.

Hobert, J., Casella, G., 1996. The Effect of Improper Priors on Gibbs Sampling in Hierarchical Linear Mixed Models. Journal of the American Statistical Association 91 (436), 1461–1473.

Huang, H. C., 1999. Estimation of sur tobit model via the mcecm algorithm. Economics Letters 64, 25–30.

Hughes, J., 1999. Mixed effects models with censored data with application to HIV RNA levels. Biometrics 55 (2), 625–629.

Jreskog, K. G., 1969. A general approach to confirmatory maximum likelihood factor analysis. Psychometrika 34 (2), 183–202.

Kamakura, W. A., Wedel, M., 2001. Exploratory tobit factor analysis for multivariate censored data. Multivariate Behavioral Research 36, 53–82.

Kass, R., Tierney, L., Kadane, J., 1989. Approximate methods for assessing influence and sensitivity in Bayesian analysis. Biometrika 76, 663–674.

Kember, D., Leung, D. Y., 1998. The dimensionality of approaches to learning: An investigation with confirmatory factor analysis on the structure of the spq and lpq. British Journal of Educational Psychology 68 (3), 395–407.

Lachos, V. H., Castro, L. M., Dey, D. K., 2013. Bayesian inference in nonlinear mixed–effects models using normal independent distributions. Computational Statistics & Data Analysis 64, 237–252.

Lopes, H. F., West, M., 2004. Bayesian model assessment in factor analysis. Statistica Sinica 14 (1), 41–68.

Matos, L., Lachos, V., Balakrishnan, N., Labra, F., 2013. Influence diagnostics in linear and nonlinear mixed-effects models with censored data. Computational Statistical & Data Analysis 57 (1), 450–464.

Muthén, B. O., 1989. Tobit factor analysis. British Journal of Mathematical and Statistical Psychology 42 (2), 241–250.

Peng, F., Dey, D. K., 1995. Bayesian analysis of outlier problems using divergence measures. The Canadian Journal of Statistics 23, 199–213.

Pilati, R., Laros, J. A., 2007. Modelos de equaes estruturais em psicologia: conceitos e aplicaes. Psicologia: Teoria e Pesquisa 23 (2), 205–216.

R Core Team, 2012. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0. URL `http://www.R-project.org/`

RTI International, 2009. Early Grade Reading Assessment Toolkit. Research Triangle Institute, USA.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., van der Linde, A., 2002. Bayesian measures of model complexity and fit 64 (4), 583–639.

Tobin, J., 1958. Estimation of relationships for limited dependent variables. Econometrica 26, 24–36.

Ullman, J. B., 2006. Structural equation modeling: Reviewing the basics and moving forward. Journal of Personality Assessment 87 (1), 35–50.

Vaida, F., Liu, L., 2009. Fast Implementation for Normal Mixed Effects Models With Censored Response. Journal of Computational and Graphical Statistics 18 (4), 797–817.

Vidal, I., Castro, L. M., 2010. Influential observations in the independent Student-t measurement error model with weak nondifferential error. Chilean Journal of Statistics 1 (2), 17–34.

Waller, N. G., Muthn, B. O., 1992. Genetic tobit factor analysis: quantitative genetic modeling with censored data. Behavior genetics 22 (3), 265–292.

Weiss, R., 1996. An approach to Bayesian sensitivity analysis. Journal of the Royal Statistical Society, Series B 58, 739–750.

Zhou, X., Liu, X., 2009. The Monte Carlo em method for estimating multivariate tobit latent variable models. Journal of Statistical Computation and Simulation 79, 1095–1107.