DENISE REIS COSTA

ESTIMAÇÃO ROBUSTA EM MODELOS DE VARIÁVEIS LATENTES PARA DADOS CENSURADOS

CAMPINAS

2013

UNICAMP

UNIVERSIDADE ESTADUAL DE CAMPINAS

INSTITUTO DE MATEMÁTICA, ESTATÍSTICA
E COMPUTAÇÃO CIENTÍFICA

DENISE REIS COSTA

ESTIMAÇÃO ROBUSTA EM MODELOS DE VARIÁVEIS LATENTES PARA DADOS CENSURADOS

Tese apresentada ao Instituto de Matemática, Estatística e Computação Científica da Universidade Estadual de Campinas como parte dos requisitos exigidos para a obtenção do título de Doutora em Estatística.

Orientador: Prof. Dr. Víctor Hugo Lachos Dávila

Este exemplar corresponde à versão final da tese defendida pela aluna Denise Reis Costa e orientada pelo Prof Dr. Víctor Hugo Lachos Dávila.

Assinatura do Orientador

Campinas

2013

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca do Instituto de Matemática, Estatística e Computação Científica
Maria Fabiana Bezerra Muller - CRB 8/6162

C823e

Costa, Denise Reis, 1985-
Estimação robusta em modelos de variáveis latentes para dados censurados / Denise Reis Costa. – Campinas, SP : [s.n.], 2013.

Orientador: Víctor Hugo Lachos Dávila.
Tese (doutorado) – Universidade Estadual de Campinas, Instituto de Matemática, Estatística e Computação Científica.

1. Variáveis latentes. 2. Estimativa de parâmetro. 3. Métodos estatísticos robustos. I. Lachos Dávila, Víctor Hugo,1973-. II. Universidade Estadual de Campinas. Instituto de Matemática, Estatística e Computação Científica. III. Título.

Informações para Biblioteca Digital

Título em outro idioma: Robust estimation in latent variable models for censored data
Palavras-chave em inglês:
Latent variables
Parameter estimation
Robust statistical methods
Área de concentração: Estatística
Titulação: Doutora em Estatística
Banca examinadora:
Víctor Hugo Lachos Dávila [Orientador]
Hildete Prisco Pinheiro
Luis Mauricio Castro Cepero
Dipankar Bandyopadhyay
Celso Rômulo Barbosa Cabral
Data de defesa: 02-12-2013
Programa de Pós-Graduação: Estatística

Tese de Doutorado defendida em 02 de dezembro de 2013 e aprovada

Pela Banca Examinadora composta pelos Profs. Drs.

____

Prof(a). Dr(a). VICTOR HUGO LACHOS DÁVILA

Hildete Pinheiro

____

Prof(a). Dr(a). HILDETE PRISCO PINHEIRO

____

Prof(a). Dr(a). LUIS MAURICIO CASTRO CEPERO

Dipankar Bandy

____

Prof(a). Dr(a). DIPANKAR BANDYOPADHYAY

Celso Rômulo B Cabral

____

Prof(a). Dr(a). CELSO RÔMULO BARBOSA CABRAL

Abstract

Latent variable models are broadly used by psychometrists, econometrists and social science researchers to model variables that cannnot be directly measured, known as constructs or random effects (Skrondal and Rabe-Hesketh, 2004). In the literature, such variables are commonly modeled with a normal distribution, but such assumption may be inadequate, especially when there are outliers. Concerned with the sensitivity of the inferences under the presence of potential outliers or data derived from heavy-tailed distributions, this thesis proposes robust inference models, using the mutivariate t-Student distribution, for two types of latent variable models: the Generalized Linear Mixed Model for correlated binary data (GLMM) and the Tobit Confirmatory Factor Analysis (TCFA) for continuous and censored data. In order to estimate the parameters of the studied models, an EM-type algorithm was proposed. This algorithm presents closed expressions on the E-step which use the two first moments of a multivariate truncated t-distribution. Moreover, we present a Bayesian approach and propose measures of influence diagnostics for censored data under the TCFA model when normality is assumed. In order to evaluate the proposed methods, simulated studies were carried out, as well as the application on real datasets.

Resumo

Modelos de variáveis latentes são amplamente utilizados por psicometristas, econometristas e pesquisadores da área de ciencias sociais para modelar variáveis que não podem ser medidas diretamente, conhecidas como construtos ou efeitos aleatórios (Skrondal e Rabe-Hesketh, 2004). Na literatura, é muito comum verificar a utilização da distribuição normal para a modelagem dessas variáveis, contudo tal suposição pode ser inadequada, especialmente na presença de valores discrepantes. Preocupados com a sensibilidade das inferências sob a presença de potenciais pontos discrepantes ou com dados provenientes de distribuições com caudas pesadas, nesta tese propomos métodos de inferência robusta, utilizando a distribuição t de Student multivariada, para dois tipos de modelos de variáveis latentes: o modelo linear generalizado misto para respostas binárias (GLMM) e o modelo de análise fatorial Tobit (TCFA) para respostas contínuas e censuradas. Para a estimação dos parâmetros dos modelos estudados, um algoritmo do tipo EM foi proposto e este apresenta expressões fechadas no passo E que utiliza os dois primeiros momentos de uma distribuição multivariada t truncada. Adicionalmente apresentamos uma abordagem via análise Bayesiana e propomos medidas de diagnóstico de influência para dados censurados sob o modelo TCFA quando a suposição de normalidade é assumida. Para avaliação dos métodos propostos, foram realizados alguns estudos simulados, além da aplicação a conjuntos de dados reais.

Sumário

Dedicatória

Dedico esta tese à minha grande amiga, Vera.
Tivemos uma jornada parecida,
compartilhamos muitos momentos juntas,
mas Deus quis levá-la mais cedo.
Obrigada por fazer parte da minha vida!

“Amigo é coisa para se guardar
Debaixo de sete chaves
Dentro do coração
Assim falava a canção que na América ouvi
Mas quem cantava chorou
Ao ver o seu amigo partir
Mas quem ficou, no pensamento voou
Com seu canto que o outro lembrou
E quem voou, no pensamento ficou
Com a lembrança que o outro cantou
Amigo é coisa para se guardar
No lado esquerdo do peito
Mesmo que o tempo e a distância digam “não”
Mesmo esquecendo a canção
O que importa é ouvir
A voz que vem do coração
Pois seja o que vier, venha o que vier
Qualquer dia, amigo, eu volto
A te encontrar
Qualquer dia, amigo, a gente vai se encontrar.”
Milton Nascimento e Fernando Brant

Agradecimentos

Imensamente grata sou por todos que torceram por mim, rezaram, colaboraram diretamente ou indiretamente para a conclusão deste trabalho. Em especial, eu gostaria de agradecer:

- A Deus, todo poderoso, que concede muitas graças em minha vida. Obrigada por me permitir concluir mais esta etapa. Obrigada pela oportunidade de conhecer pessoas maravilhosas nesta jornada e pelo tanto que aprendi.
- Ao meu orientador Prof. Víctor H. Lachos por sua dedicação, incentivo, confiança e imensa paciência. Seus ensinamentos contribuíram de forma significativa para meu crescimento pessoal e formação profissional.
- Ao meus pais, grandes exemplos na minha vida. Obrigada por todo amor, compreensão e saudade nesses anos longe.
- A minhas irmãs, cunhados e sobrinhos que sempre me apoiaram e torceram por mim.
- Ao meu esposo, meu grande companheiro e amigo. Agradeço por ter aguentado todo mix de emoções (alegria, tristeza, euforia, raiva) ocorrido durante esses anos do doutorado. Te amo muito!
- À família do meu esposo, pelo carinho incondicional.
- Aos professores Dr. Marcos Prates da UFMG, Dr. Jorge Bazán da USP/São Carlos, Dr. Caio Azevedo da Unicamp e membros da comissão examinadora por sua disponibilidade e contribuição.
- Aos amigos que fiz em Campinas, aos que tinham ficado em Brasília e aos que conheci em outros cantos, agradeço o incentivo e todo suporte emocional e afetivo.
- Aos professores do Programa de Pós-Graduação em Estatística e aos funcionários do IMECC, obrigada pelo profissionalismo e atenção.
- À Capes pela bolsa concedida em parte do programa.

A todos, meus sinceros agradecimentos!

Capítulo 1

Introdução

1.1 Aspectos gerais

Modelos de variáveis latentes têm sido amplamente utilizados em áreas tais como medicina, economia, psicologia, e marketing para obter estimativas de variáveis que não podem ser medidas diretamente. Inicialmente esses modelos foram vistos como uma área obscura da estatística, primariamente confinada à psicometria. Contudo, eles vêm ganhando espaço na literatura por permitir gerar distribuições multivariadas flexíveis e combinar a informação acerca das unidades individuais a partir de diferentes fontes (Skrondal and Rabe-Hesketh, 2004).

Usualmente, esses modelos envolvem dois conjuntos de variáveis: uma que é observável (variável resposta ou manifesta) e outra subjacente (variável latente). Nenhuma restrição é feita a respeito da variável resposta que pode ser discreta, contínua, categórica ou mista (Bartholomew, 1984). Contudo, as variáveis latentes são consideradas, em geral, contínuas. Em muitos casos, essas variáveis não são diretamente reconhecidas, possivelmente porque variáveis latentes apresentam diferentes denominações, como, por exemplo, efeitos aleatórios, fatores comuns e classes latentes. Elas representam importantes fenômenos de análise, tais como construção de construtos, captura de heterogeneidade não observada, resolução de problemas de dados faltantes ou de respostas latentes adjacentes de uma variável categórica.

Segundo Skrondal and Rabe-Hesketh (2004), as abordagens usuais para a modelagem dessas variáveis são: modelos multiníveis, modelos longitudinais, modelos da Teoria de Resposta ao Item (TRI) e modelos de Análise Fatorial. A inferência estatística para modelagem de variáveis respostas contínuas assumindo normalidade dos dados está bem desenvolvida e implementada em vários *softwares* estatísticos, tais como procedimento MIXED no SAS e LISREL para análise fatorial (An and Bentler, 2012). Todavia, inferências para variáveis respostas categóricas ou censuradas apresentam-se ainda em fase de desenvolvimento.

Ressalta-se, no entanto, que encaixam-se na definição de dados censurados a que se refere o presente trabalho aquelas observações que atingem um valor máximo ou mínimo (efeitos *ceiling* e *floor*, respectivamente) devido a alguma característica associada à natureza dos dados (Waller and Muthén, 1992). A Figura 1.1 exemplifica a configuração de uma análise de dados multivariados censurados no valor zero. Eles serão melhor descritos

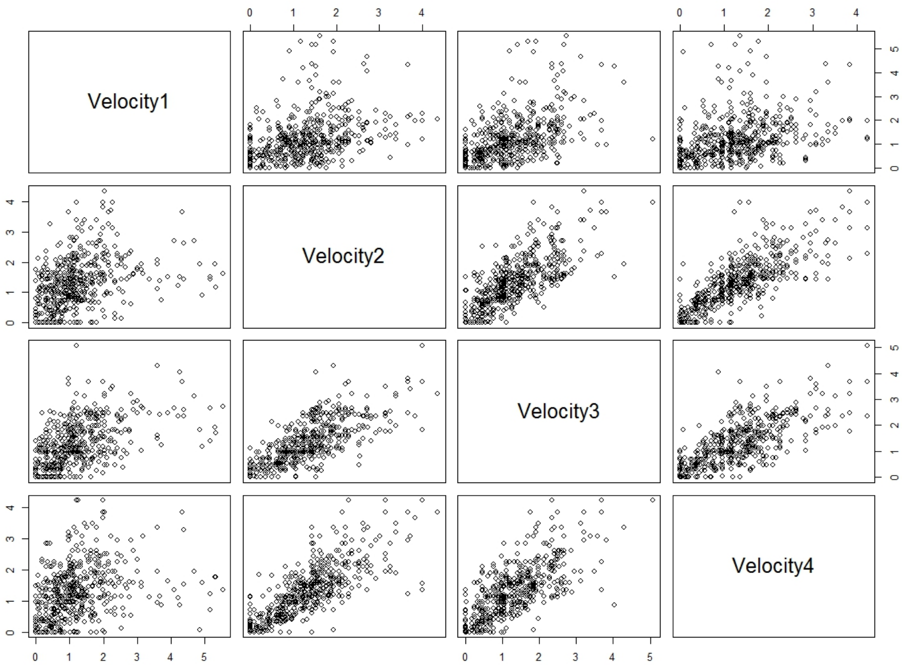

Figura 1.1: Gráficos de dispersão das medidas de velocidade - Banco de dados EGRA.

no Capítulo 3, mas podemos observar de antemão que há alta proporção de casos atingindo o valor mínimo da escala (efeito *floor*). Incluir essa informação na modelagem é o tema desta tese.

Note ainda que variáveis binárias podem, implicitamente, ser consideradas como variáveis censuradas atingindo um limite mínimo em zero e máximo em 1. Segundo Greene (2005), um exemplo clássico desse tipo de censura é a interpretação da regressão latente de uma variável binária em que, por exemplo, a variável latente x^* representa a preferência de um eleitor em um sistema bipartidário e x (variável observável) denota em qual dentre os dois partidos o eleitor votou. Dessa forma, um modelo de variável latente pode ser construído da seguinte forma:

$$X = \begin{cases} 0, & \text{se } X^* \leq 0 \\ 1, & \text{se } X^* > 0, \end{cases} \tag{1.1.1}$$

em que X^* segue alguma distribuição conhecida. Modelos como o probito, introduzido por Bliss (1935), que atribuem normalidade à variável latente X^* são corriqueiramente utilizados na literatura para se ajustar variáveis do tipo (1.1.1). Por outro lado, quando a variável observável é contínua, não negativa e censurada em zero, Tobin (1958) introduziu uma classe de modelos denominados modelos Tobit. A Figura 1.2 ilustra uma variável associada a esse modelo. Valores menores ou iguais a x^* são censuradas nesse limite.

Dentre a ampla variedade de modelos que poderiam ser utilizados para se trabalhar com essas variáveis, nesta tese o foco se concentra em dois: o modelo linear generalizado misto para respostas binárias (GLMM) e o modelo de análise fatorial Tobit (TCFA) com covariáveis. Enquanto o primeiro se preocupa com a modelagem de dados do tipo (1.1.1) advindos de dados longitudinais ou de medidas repetidas, o segundo modela a correlação entre variáveis contínuas censuradas tal como apresentado na Figura 1.1.

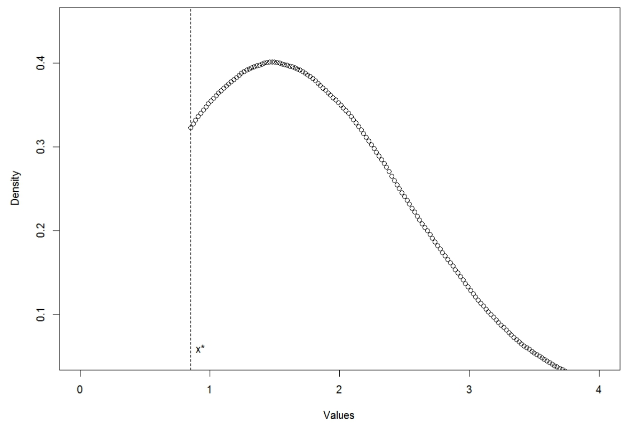

Figura 1.2: Exemplo de uma variável normalmente distribuída com censura em x^*.

Por entender que, em muitas situações, inferências estatísticas sob normalidade são impróprias, como, por exemplo, na presença de *outliers*, a principal contribuição metododológica desse trabalho é o de propor o desenvolvimento de métodos robustos de estimação em modelos de variáveis latentes. Devido a sua complexidade, pouco se tem observado na literatura de modelos GLMM e de análise fatorial, especialmente, sobre modelos alternativos ao modelo normal que preservam a estrutura simétrica e que permitam reduzir a influência de observações discrepantes. A substituição da distribuição normal por distribuições de caudas pesadas, tal como a distribuição t de Student, tem sido apresentada recentemente por Li et al. (2007) e Zhou and Tan (2010). Diferentemente da metodologia utilizada por esses autores, propomos novos mecanismos de análise, sob perspectiva clássica, utilizando distribuições multivariadas t de Student usual e t de Student truncada. Para tanto, a seguir será feita uma breve descrição dessas distribuições.

1.1.1 Distribuições multivariada t de Student usual e truncada

Uma variável aleatória com distribuição t de Student p-variada com vetor de locação $\boldsymbol{\mu}$, matriz de escala $\boldsymbol{\Sigma}$ e ν graus de liberdade, denotada por $t_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$, apresenta a seguinte representação (Lachos et al., 2013a):

$$\mathbf{Y} = \boldsymbol{\mu} + U^{-1/2}\mathbf{Z}, \;\; \mathbf{Z} \sim N_p(\mathbf{0}, \boldsymbol{\Sigma}), \; U \sim \text{Gamma}(\nu/2, \nu/2), \tag{1.1.2}$$

em que $\mathbf{Z}$ e U são independentes e Gamma(a, b) denota uma distribuição Gama com média a/b. A densidade de probabilidade de $\mathbf{Y}$ é dada por:

$$t_p(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu) = \frac{\Gamma(\frac{p+\nu}{2})}{\Gamma(\frac{\nu}{2})\pi^{p/2}} \nu^{-p/2} |\boldsymbol{\Sigma}|^{-1/2} \left(1 + \frac{\delta}{\nu}\right)^{-(p+\nu)/2},$$

em que $\Gamma(.)$ é a função gamma padrão e $\delta = (\mathbf{y} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{y} - \boldsymbol{\mu})$, a distância de Mahalanobis. A função de distribuição acumulada é denotada por $T_p(.|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$. Se $\nu > 1$, $\boldsymbol{\mu}$ é a média de $\mathbf{Y}$, e se $\nu > 2$, $\nu(\nu - 2)^{-1}\boldsymbol{\Sigma}$ será sua matriz de covariância. Quando ν tende ao infinito, U converge para 1 com probabilidade igual a um, e então $\mathbf{Y}$ tende marginalmente para uma normal multivariada com média $\boldsymbol{\mu}$ e matriz de covariância $\boldsymbol{\Sigma}$.

Denotando $Tt_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu; \mathbb{A})$ uma distribuição t de Student truncada de dimensão p para $t_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$ em um hiperplano truncado à direita representado por

$$\mathbb{A} = \{\mathbf{x} = (x_1, \ldots, x_p)^\top | x_1 \leq a_1, \ldots, x_p \leq a_p\}. \tag{1.1.3}$$

Especificamente, pode-se dizer que um vetor p-dimensional $\mathbf{X}$ segue distribuição $Tt_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu; \mathbb{A})$ se sua densidade é dada por $f(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu; \mathbb{A}) = \dfrac{t_p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)}{T_p(\mathbf{a}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)}\mathbb{I}_{\mathbb{A}}(\mathbf{x})$, em que $\mathbf{a} = (a_1, \ldots, a_p)^\top$ e $\mathbb{I}_{\mathbb{A}}(\mathbf{x})$ é uma função indicadora cujos valores são iguais a um se $\mathbf{x} \in \mathbb{A}$ e zero, caso contrário.

Utilizando resultados de Matos et al. (2013b), os momentos de uma distribuição multivariada t truncada à direita serão úteis para o desenvolvimento da teoria proposta neste trabalho e serão descritos na seguinte Proposição.

Proposição 1.1.1. Se $\mathbf{X} \sim Tt_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu; \mathbb{A})$ com $\mathbb{A}$ como definida em (1.1.3), então o k-ésimo momento de $\mathbf{X}$, $k = 0, 1, 2$, é dado por:

$$\begin{aligned} E\left\{\left(\frac{\nu+p}{\nu+\delta}\right)^r \mathbf{X}^{(k)}\right\} &= c_p(\nu, r)\frac{T_p(\mathbf{a}|\boldsymbol{\mu}, \boldsymbol{\Sigma}^*, \nu+2r)}{T_p(\mathbf{a}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)}E_{\mathbf{W}}\{\mathbf{W}^{(k)}\}, \\ \mathbf{W} &\sim Tt_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}^*, \nu+2r; \mathbb{A}), \end{aligned}$$

em que $c_p(\nu, r) = \left(\dfrac{\nu+p}{\nu}\right)^r \left(\dfrac{\Gamma((p+\nu)/2)\Gamma((\nu+2r)/2)}{\Gamma(\nu/2)\Gamma((p+\nu+2r)/2)}\right)$, $\delta = (\mathbf{X} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{X} - \boldsymbol{\mu})$, $\mathbf{a} = (a_1, \ldots, a_p)^\top$, $\boldsymbol{\Sigma}^* = \dfrac{\nu}{\nu+2r}\boldsymbol{\Sigma}$, $\mathbf{X}^{(0)} = 1$, $\mathbf{X}^{(1)} = \mathbf{X}$, $\mathbf{X}^{(2)} = \mathbf{X}\mathbf{X}^\top$ e $\nu + 2r > 0$.

Diferentemente de Tan et al. (2007), Meza et al. (2009), An and Bentler (2012) que propuseram métodos que utilizam o algoritmo EM (Dempster et al., 1977) conjuntamente à amostragem via Monte Carlo para se obter estimativas de máxima verossimilhança dos parâmetros de modelo de variáveis latentes, neste trabalho desenvolvemos uma nova abordagem para esse problema, via algoritmo EM, em que há expressões fechadas para implementação do algoritmo, graças a fórmulas para $E[\mathbf{W}]$ e $E[\mathbf{W}\mathbf{W}^\top]$, que têm sido recentemente desenvolvidas por Ho et al. (2012) e podem ser facilmente obtidas utilizando os pacotes *pmvt*() e *mvtnorm* (Genz et al., 2008) disponibilizados no *software* R.

A seguir, será apresentada uma breve descrição do algoritmo EM que será utilizado nos capítulos seguintes. Também será apresentada uma breve descrição do amostrador de Gibbs utilizado para se fazer inferências sob perspectiva Bayesiana para o modelo TCFA.

1.1.2 Algoritmo EM

Para se obter estimativas de máxima verossimilhança (MV) dos modelos propostos, o algoritmo EM será utilizado. O algoritmo EM (Dempster et al., 1977) consiste em um processo iterativo para se obter estimativas MV de um vetor de parâmetros na presença de variáveis não observáveis (variáveis latentes). Esse algoritmo se torna bastante útil, pois permite utilizar uma formulação de dados aumentados (também chamados de dados completos) que simplifica o processo de estimação dos parâmetros do modelo. Para melhor entendimento, esse procedimento será descrito brevemente.

Considere $\mathbf{Y}_{obs}$ os dados observados, $\mathbf{Y}_{mis}$ os dados faltantes e $\mathbf{Y}_{com} = \{\mathbf{Y}_{obs}, \mathbf{Y}_{mis}\}$ os dados completos. Denote também $L(\boldsymbol{\theta}|\mathbf{Y}_{com})$ a função de verossimilhança dos dados completos, $\ell(\boldsymbol{\theta}|\mathbf{Y}_{com}) = \log L(\boldsymbol{\theta}|\mathbf{Y}_{com})$, a função log-verossimilhança, em que $\boldsymbol{\theta} \in \Theta$. Um importante elemento para implementação do algoritmo EM é dado pela esperança da log-verossimilhança dos dados completos, $Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k)})$, dada por:

$$Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k)}) = E[\ell(\boldsymbol{\theta}|\mathbf{Y}_{com})|\mathbf{Y}, \widehat{\boldsymbol{\theta}}^{(k)}].$$

Um resumo sobre os passos do processo iterativo são apresentados a seguir:

Passo E Calcula-se $Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k)})$ como função de $\boldsymbol{\theta}$;

Passo M Encontra-se $\boldsymbol{\theta}^{(k+1)}$ tal que $Q(\boldsymbol{\theta}^{(k+1)}|\widehat{\boldsymbol{\theta}}^{(k)}) = max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k)})$

Cada iteração do algoritmo EM incrementa o logaritmo da função de verossimilhança observada $\ell(\boldsymbol{\theta}|\mathbf{Y}_{obs})$ de modo que $\ell(\boldsymbol{\theta}^{(k)}|\mathbf{Y}_{obs}) \leq \ell(\boldsymbol{\theta}^{(k+1)}|\mathbf{Y}_{obs})$, convergindo, assim, para um ponto de máximo local ou global da função de verossimilhança avaliada.

Quando a maximização simultânea de todas as componentes do vetor de parâmetros é de difícil implementação, uma alternativa muito utilizada é a de se substituir este por uma sequência de passos de maximização restrita. Esse procedimento leva a uma extensão simples do algoritmo EM, denominado algoritmo de maximização condicional (ECM) (Meng and van Dyk, 1998).

1.1.3 Amostrador de Gibbs

Neste trabalho propomos também uma abordagem via análise Bayesiana para o modelo TCFA quando a distribuição Normal é assumida. Para tanto, uma breve descrição do amostrador de Gibbs será realizada.

O amostrador de Gibbs, introduzido por Geman and Geman (1984), é uma das técnicas de Monte Carlo via cadeias de Markov (MCMC) muito utilizadas para se fazer inferências acerca dos parâmetros do modelo a partir de amostras geradas da distribuição *a posteriori*. O método consiste no uso de distribuições condicionais completas associadas aos parâmetros como distribuições de transição da cadeia de Markov.

Assumindo que a distribuição *a posteriori* é dada por $\pi(\boldsymbol{\theta})$, com $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_I)^\top$, a distribuição condicional completa do parâmetro θ_i é dada por $\pi_i(\boldsymbol{\theta}_i) = \pi(\boldsymbol{\theta}_i)/\pi(\boldsymbol{\theta}_{(-i)})$, $i = (1, \ldots, I)$, em que $\boldsymbol{\theta}_{(-i)}$ corresponde ao vetor $\boldsymbol{\theta}$ sem a i-ésima componente. Sendo assim, o algoritmo do amostrador de Gibbs se sintetiza da seguinte maneira:

1. Inicialize o contador de iterações $k = 1$ e o conjunto arbitrário de valores iniciais $\boldsymbol{\theta}^{(0)} = (\theta_1^{(0)}, \ldots, \theta_I^{(0)})^\top$

2. Obtenha um novo valor do vetor de parâmetros $\boldsymbol{\theta}^{(k)} = (\theta_1^{(k)}, \ldots, \theta_I^{(k)})^\top$ através de gerações sucessivas de valores amostrados de

$$\begin{aligned}
\theta_1^{(k)} &\sim \pi_i(\theta_1|\theta_2^{(k-1)}, \ldots, \theta_I^{(k-1)}) \\
\theta_2^{(k)} &\sim \pi_i(\theta_2|\theta_1^{(k)}, \theta_3^{(k-1)}, \ldots, \theta_I^{(k-1)}) \\
&\cdots \\
\theta_I^{(k)} &\sim \pi_i(\theta_I|\theta_1^{(k)}, \ldots, \theta_{I-1}^{(k)})
\end{aligned}$$

3. Incremente o contador k para k+1 e refaça o item 2 até que algum critério de convergência seja alcançado.

1.2 Objetivos e organização do trabalho

Nesta tese, pretendemos fazer um estudo de inferência estatística robusta para o modelo linear generalizado misto para respostas binárias (GLMM) e para o modelo de análise fatorial Tobit (TCFA). O objetivo principal deste trabalho é o de propor, apresentar e desenvolver métodos de estimação utilizando a distribuição multivariada t de Student para esses modelos. Para tanto, foram estruturados os seguintes capítulos como objetivos específicos:

Capítulo 1: Nesse capítulo foram apresentados os aspectos gerais que motivaram o estudo. Uma breve descrição da distribuição t de Student usual e truncada foi apresentada, bem como a apresentação dos algoritmos EM e do amostrador de Gibbs que serão utilizados ao longo do texto.

Capítulo 2: Discutiremos uma generalização do trabalho de Tan et al. (2010) para estimação robusta de parâmetros em modelos lineares generalizados mistos (GLMM) para dados binários correlacionados. Nesse capítulo apresentamos a metodologia proposta bem como a aplicação dos métodos estudados para dados simulados e dados reais.

Capítulo 3: Nesse capítulo serão apresentados os métodos desenvolvidos para estimação por máxima verossimilhança, via algoritmo EM, e pela análise Bayesiana para inferência no modelo de análise fatorial confirmatória Tobit com covariáveis. Derivamos também medidas de diagnóstico para detecção de observações discrepantes e desenvolvemos também um estudo de estimação para esse modelo utilizando a distribuição t de Student.

Capítulo 4: Serão apresentadas algumas considerações finais, além de apresentar sugestões para pesquisa futura.

Capítulo2

Modelos Lineares Generalizados Mistos para dados binários correlacionados

Neste capítulo apresentaremos um resumo dos resultados apresentados no artigo intitulado "Generalized linear mixed models for correlated binary data with t-link" descrito no Apêndice A. Este trabalho foi aceito para publicação na revista *Statistics and Computing* e contou com a colaboração do Prof. Dr. Marcos Prates da Universidade Federal de Minas Gerais (UFMG).

2.1 Introdução

Modelos Lineares Generalizados Mistos (GLMM) são extensões naturais de modelos lineares generalizados para análise de dados coletados de diferentes *clusters* ou de estudos longitudionais (Breslow and Clayton, 1993). Em geral, as características populacionais são modeladas como efeitos fixos e as variações individuais como efeitos aleatórios. Uma popular classe desses modelos é a de função de ligação probito-normal para análise de dados binários ou em escala ordinal. Apesar dos métodos baseados na verossimilhança para modelos GLMM de respostas contínuas serem bem desenvolvidos (Meng and van Dyk, 1998), a estimação de Máxima Verossimilhança de modelos GLMM para dados binários correlacionados ainda apresenta desafios devido à complexidade de sua função de verossimilhança. Dessa maneira, observa-se na literatura crescente interesse no desenvolvimento de métodos via máxima verossimilhança para essa classe de modelos. Por exemplo, McCulloch (1994) propôs um algoritmo Monte Carlo EM (MCEM) utilizando o amostrador de Gibbs no passo E para se modelar respostas binárias com função de ligação probito. Mais tarde, esse mesmo autor implementou passos de Metropolis–Hastings no algoritmo MCEM para se ajustar uma classe maior de modelos GLMM. Tan et al. (2007), por sua vez, propuseram uma abordagem não iterativa via função de importância (*importance sampling*) em que o algoritmo MCEM utiliza os dois primeiros momentos de uma distribuição Normal multivariada truncada. Já Meza et al. (2009) utilizaram uma aproximação estocástica do algoritmo EM (SAEM), apresentado por Delyon et al. (1999), para se obter estimativas via método de Máxima Verossimilhança. Lee and Nelder (2006), no entanto, desenvolveram um procedimento com o uso de uma verossimilhança do tipo hierárquica (ou verossimilhança-h) para se fazer inferência para esses modelos.

Apesar do avanço do desenvolvimento de métodos para estimação via máxima verossimilhança de modelos GLMM para dados binários correlacionados, observa-se ainda que grande parte dos métodos desenvolvidos utilizam alguma combinação de métodos de Monte Carlo com o algoritmo EM, que demandam grande esforço computacional. Além disso, a maioria dos trabalhos nesta área utilizam como função de ligação a probito-normal, que nem sempre se ajusta adequadamente aos dados. No entanto, funções de ligação mal especificadas podem levar a um significativo vício nas estimativas que comprometem seriamente as inferências (Czado and Santner, 1992).

Uma maneira de se prevenir a má especificação da função de ligação é utilizar uma classe de funções paramétricas mais gerais em que as funções probito e logito são casos particulares. Seguindo Liu (2004), neste capítulo propomos uma modelagem paramétrica robusta para o modelo GLMM para dados binários tendo por base a função de ligação t, denominado t-GLMM. Uma abordagem via máxima verossimilhança foi desenvolvida, incluindo a implementação de um algoritmo ECM exato. Sob nossa abordagem, as funções de ligação probito e logito podem sem consideradas como casos especiais. Tal como apresentado por Matos et al. (2013b), será mostrado que os passos E do algoritmo EM se reduz ao cálculo dos dois primeiros momentos de uma distribuição multivariada t de Student truncada. As fórmulas para se obter esses momentos são derivadas do trabalho de Ho et al. (2012) (eq. 12 e 13) que dependem da função de distribuição acumulada de uma distribuição multivariada t. Para se obter essas estimativas, o pacote `mvtnorm` (Genz et al., 2008) do R (R Core Team, 2012) foi utilizado. Por meio dessa abordagem, a função de verossimilhança pode ser facilmente calculada tornando, assim, possível sua utilização para análise da convergencia do algoritmo, bem como para se calcular medidas de comparação de modelos, tais como o critério de informação de Akaike (AIC) (Akaike, 1974), o critério de informação Bayesiano (BIC) (Schwarz, 1978) e o teste da razão de verossimilhança (LRT). Por meio dessas medidas, os resultados numéricos deste trabalho indicam que a função de ligação t de Student apresenta melhor desempenho com relação ao probito no modelo GLMM para dados binários em diferentes cenários.

Este capítulo está estruturado da seguinte maneira: na Seção 2.2, o modelo de ligação probito para o modelo GLMM (probito-GLMM) para dados binários é formulado. O modelo t-GLMM é apresentado na Seção 2.3. O novo algoritmo do tipo EM é desenvolvido na Seção 2.4, bem a descrição de alguns métodos inferencias desenvolvidos (estimação dos erros-padrão e dos efeitos aleatórios). Na Seção 2.5, será apresentado um resumo dos resultados encontrados nos estudos simulados e da aplicação a um banco de dados reais.

2.2 Modelo probito-GLMM para dados binários

Sejam Y_{ij} a variável resposta que assume os valores 0 ou 1 da $j-$ésima medida e $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{in_i})^\top$ uma coleção de respostas para o indivíduo i, em que $i = 1, \ldots, m$ e $j = 1, \ldots, n_i$. O modelo linear generalizado misto com função de ligação probito (McCulloch, 1994) assume que, dadas as variáveis latentes (efeitos aleatórios) $\mathbf{b}_i$, as respostas $\{Y_{ij}\}_{j=1}^{n_i}$ são condicionalmente independentes com probabilidade:

$$Pr(Y_{ij} = 1|\mathbf{b}_i) = \Phi(\mu_{ij}), \;\; \mu_{ij} = \mathbf{x}_{ij}^{\top}\boldsymbol{\beta} + \mathbf{w}_{ij}^{\top}\mathbf{b}_i, \mathbf{b}_i \sim N_q(\mathbf{0}, \mathbf{D}), \tag{2.2.1}$$

em que $\Phi(\cdot)$ denota a função de distribuição acumulada da $N(0,1)$, $\mathbf{x}_{ij}^{\top} = (x_{ij1}, \ldots, x_{ijp})^{\top}$ e $\mathbf{w}_{ij}^{\top} = (w_{ij1}, \ldots, w_{ijq})^{\top}$ são covariáveis, $\boldsymbol{\beta}$ é um vetor $(p \times 1)$ dos efeitos fixos, $\{\mathbf{b}_i\}_{i=1}^{m}$ é um vetor $(q \times 1)$ das variáveis latentes (efeitos aleatórios), $\mathbf{D}$ é uma matriz desconhecida $(q \times q)$ relacionada com a estrutura de correlação de $\mathbf{Y}_i$. Esse modelo pode ser alternativamente escrito em termos de uma variável contínua não-observável $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{in_i})^{\top}$, tal que

$$Y_{ij} = \mathbb{I}_{(0,\infty)}(Z_{ij}), \qquad \mathbf{Z}_i|\mathbf{b}_i \sim N_{ni}(\boldsymbol{\mu}_i, \mathbf{I}_{n_i}), \qquad \mathbf{b}_i \sim N_q(\mathbf{0}, \mathbf{D}),$$

em que $\mathbf{I}_{ni}$ denota uma matriz identidade $(n_i \times n_i)$ e $\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{in_i})^{\top} = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{W}_i\mathbf{b}_i$, com $\mathbf{X}_i = (\mathbf{x}_{i1} \ldots, \mathbf{x}_{in_i})^{\top}$, $\mathbf{W}_i = (\mathbf{w}_{i1} \ldots, \mathbf{w}_{in_i})^{\top}$ sendo matrizes de dimensão $(n_i \times p)$ e $(n_i \times q)$, respectivamente.

Sob normalidade, vários métodos têm sido propostos na literatura estatística que eficientemente estima via método de máxima verossimilhança os parâmetros associados à matriz de variância $(\mathbf{D})$, bem como os parâmetros associados aos efeitos fixos $(\boldsymbol{\beta})$. Por exemplo, um algoritmo Monte Carlo EM (MCEM) com amostrador de Gibbs a cada passo E foi proposto por McCulloch (1994). Para ajustar modelos mais gerais, McCulloch (1997) utilizou um algoritmo com passos de Metropolis-Hastings a cada etapa E do MCEM. Um procedimento aproximado que se baseia na verossimilhança h foi proposto por Lee and Nelder (2006). Mais tarde, Tan et al. (2007) desenvolveram uma abordagem não iterativa via amostragem por importância que se baseia na fórmula inversa de Bayes (IBF), em que os dois primeiros momentos de uma distribuição normal multivariada truncada podem ser avaliados e associados ao algoritmo MCEM. Meza et al. (2009), por sua vez, propuseram o uso de um algoritmo EM com aproximação estocástica (SAEM). Alternativamente, inferências Bayesianas podem ser encontradas na literatura que utilizam cadeias de Markov Monte Carlo (MCMC) e implementadas via amostrador de Gibbs (Albert and Chib, 1993). Na próxima seção, será apresentada uma proposta robusta de análise de modelo GLMM com função de ligação t de Student.

2.3 Modelo t-GLMM para dados binários

Pinheiro et al. (2001) propuseram um modelo hierárquico linear de efeitos mistos robusto em que os efeitos aleatórios e os erros seguem uma distribuição multivariada t de Student. Na formulação deles, assumem-se que as distribuições de misturas para as duas fontes de variabilidade no modelo tem o mesmo formato e compartilham dos mesmos parâmetros. Nossa formulação também está alinhada ao trabalho de Pinheiro et al. (2001) (veja também Matos et al., 2013b) por considerar uma generalização do modelo clássico probito-GLMM para dados binários (2.2.1), segundo as seguintes características:

$$\begin{aligned} Y_{ij} = \mathbb{I}_{(0,\infty)}(Z_{ij}), \;\mathbf{Z}_i|\mathbf{b}_i, U_i = u_i &\sim N_{ni}(\boldsymbol{\mu}_i, u_i^{-1}\mathbf{I}_{n_i}), \\ \mathbf{b}_i|U_i = u_i &\sim N_q(\mathbf{0}, u_i^{-1}\mathbf{D}), \\ U_i &\sim \text{Gamma}(\nu/2, \nu/2). \end{aligned} \tag{2.3.1}$$

Utilizando o Lema 1 apresentado no Apêndice A, o modelo definido em (2.3.1) é equivalente à seguinte representação:

$$Pr(Y_{ij} = 1|\mathbf{b}_i) = T_1(\mu_{ij}|0, 1, \nu), \quad \mathbf{b}_i \sim t_q(\mathbf{0}, \mathbf{D}, \nu),$$

em que $\mu_{ij} = \mathbf{x}_{ij}^\top \boldsymbol{\beta} + \mathbf{w}_{ij}^\top \mathbf{b}_i$.

Inferências clássicas para o vetor de parâmetros $\boldsymbol{\theta}$ se baseiam na função de verosimilhança de $\boldsymbol{\theta}$ condicionada as observações $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)$, que nesse caso é dada por

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{m} \int t_q(\mathbf{b}_i|\mathbf{0}, \mathbf{D}, \nu)\psi_i^t(\boldsymbol{\beta}, \mathbf{b}_i)d\mathbf{b}_i, \tag{2.3.2}$$

em que $t_q(.|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$ foi definida no Capítulo 1 e

$$\psi_i^t(\boldsymbol{\beta}, \mathbf{b}_i) = \prod_{j=1}^{n_i} [T_1(\mu_{ij}|0, 1, \nu)]^{y_{ij}} [1 - T_1(\mu_{ij}|0, 1, \nu)]^{1-y_{ij}}. \tag{2.3.3}$$

Vale notar que a função de verossimilhança (2.3.2) não apresenta uma expressão analítica fechada uma vez que a função de ligação do modelo é não linear no efeito aleatório. Para se obter estimativas de máxima verossimilhança dos parâmetros associados ao modelo, um algoritmo do tipo EM é proposto neste trabalho. Esse algoritmo baseia-se nas fórmulas da média e variância de uma distribuição multivariada t de Student truncada, que pode ser calculada analiticamente.

Diversos autores têm sugerido a estimação paramétrica do parâmetro ν (veja, por exemplo, Lange and Sinsheimer, 1993; Jamshidian, 1999) tendo por base o algoritmo EM e suas variantes. Fernandez and Steel (1999), por sua vez, discutiram os problemas potenciais que surgem com a estimação dos graus de liberdade, em particular para a distribuição t de Student. Isso acontece devido à ilimitação da função de verossimilhança perto do limite do espaço dos parâmetros e, portanto, o método de estimação MV desenvolvido por Lange and Sinsheimer (1993) é questionável. Essa metodologia não fornece informação suficiente se essas estimativas correspondem a pontos de máximos locais ou globais. Lucas (1997) notou que somente sob a fixação dos graus de liberdade a estimação dos parâmetros torna-se robusta a observações extremas. A alternativa plausível (e simples) adotada neste trabalho foi a de se assumir que o parâmetro ν associado às variáveis de mistura U é conhecido. Recentes trabalhos no contexto de distribuições elípticas também têm considerado o parâmetro ν como conhecido. Veja, por exemplo, Meza et al. (2009) e Matos et al. (2013b).

2.4 O algoritmo EM

Nesta seção, serão apresentados os passos E e M do algoritmo EM proposto, em que são demonstrados como se utilizar fórmulas exatas no passo E em vez de se fazer uso de métodos de amostragem via Monte Carlo (MC). Para o modelo probito, um algoritmo EM foi proposto por McCulloch (1994), e recentemente melhorado graças aos trabalhos realizados por Meza et al. (2009) e Tan et al. (2007). Seguindo a notação deste último, vamos tratar $\mathbf{b} = \{\mathbf{b}_i\}_{i=1}^m$, $\mathbf{Z} = \{\mathbf{Z}_i\}_{i=1}^m$ e $\mathbf{U} = \{U_i\}_{i=1}^m$ como dados faltantes e $\mathbf{Y} = \{\mathbf{Y}_i\}_{i=1}^m$ como os dados observados.

Então, a função de densidade conjunta para os dados completos $\mathbf{Y}_{com} = \{\mathbf{Y}, \mathbf{Z}, \mathbf{b}, \mathbf{U}\}$ será

$$L(\boldsymbol{\theta}|\mathbf{Y}_{com}) = \prod_{i=1}^{m} \left[\phi_q(\mathbf{b}_i|\mathbf{0}, u_i^{-1}\mathbf{D})\phi_{n_i}(\mathbf{Z}_i|\boldsymbol{\mu}_i, u_i^{-1}\mathbf{I}_{n_i})G(u_i|\nu/2, \nu/2)\right].$$

O objetivo do passo M do algoritmo EM é obter estimativas de $\boldsymbol{\theta} = (\boldsymbol{\beta}, \mathbf{D})$ via máxima verossimilhança dos dados completos pela otimização da função esperada condicional do log da verossimilhança dos dados completos $\ell(\boldsymbol{\theta}|\mathbf{Y}_{com}) = \log L(\boldsymbol{\theta}|\mathbf{Y}_{com})$ dadas as observações $\mathbf{Y}$ e o valor atual de $\boldsymbol{\theta}^{(k)}$. Essa função é denotada por

$$\begin{aligned} Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k)}) &= E[\ell(\boldsymbol{\theta}|\mathbf{Y}_{com})|\mathbf{Y}, \widehat{\boldsymbol{\theta}}^{(k)}] = C - \frac{m}{2} log|\mathbf{D}| - \frac{1}{2}\sum_{i=1}^{m} tr\left(\mathbf{D}^{-1}E[U_i\mathbf{b}_i\mathbf{b}_i^\top|\mathbf{Y}_i]\right) \\ &+ \sum_{i=1}^{m}\left(\boldsymbol{\beta}^\top\mathbf{X}_i^\top\mathbf{X}_i\boldsymbol{\beta}E[U_i|\mathbf{Y}_i]\right) - 2\sum_{i=1}^{m}\left[\boldsymbol{\beta}^\top\mathbf{X}_i^\top\left(E[U_i\mathbf{Z}_i|\mathbf{Y}_i] - \mathbf{W}_iE[U_i\mathbf{b}_i|\mathbf{Y}_i]\right)\right] \end{aligned}$$

em que C é uma constante que independe dos parâmetros do modelo. Dessa forma, no passo M obtém-se estimativas atualizadas de $\boldsymbol{\beta}$ e $\mathbf{D}$ por meio das seguintes expressões fechadas:

$$\widehat{\boldsymbol{\beta}} = (\sum_{i=1}^{m} E[U_i|\mathbf{Y}_i]\mathbf{X}_i^\top\mathbf{X}_i)^{-1}\sum_{i=1}^{m}(\mathbf{X}_i^\top E[U_i\mathbf{Z}_i|\mathbf{Y}_i] - \mathbf{W}_iE[U_i\mathbf{b}_i|\mathbf{Y}_i]) \text{ e} \tag{2.4.1}$$

$$\widehat{\mathbf{D}} = \frac{1}{m}\sum_{i=1}^{m} E[U_i\mathbf{b}_i\mathbf{b}_i^\top|\mathbf{Y}_i], \tag{2.4.2}$$

em que $E[U_i|\mathbf{Y}_i]$, $E[U_i\mathbf{Z}_i|\mathbf{Y}_i]$, $E[U_i\mathbf{b}_i|\mathbf{Y}_i]$ e $E[U_i\mathbf{b}_i\mathbf{b}_i^\top|\mathbf{Y}_i]$ são os valores esperados de $(U_i, \mathbf{b}_i, \mathbf{Z}_i)$ condicional a $\mathbf{Y}_i$, levando em consideração o atual valor das estimativas dos parâmetros $\boldsymbol{\theta}^{(k)} = (\boldsymbol{\beta}^{(k)}, \mathbf{D}^{(k)})$. Logo, pode ser mostrado (veja o Apêndice A) que

$$\begin{aligned} E[U_i|\mathbf{Y}_i] &= \bar{\mathbf{Z}}_i^0,\ E[U_i\mathbf{Z}_i|\mathbf{Y}_i] = \bar{\mathbf{Z}}_i^1, \\ E[U_i\mathbf{b}_i|\mathbf{Y}_i] &= \boldsymbol{\Delta}_i(\bar{\mathbf{Z}}_i^1 - \bar{\mathbf{Z}}_i^0\mathbf{X}_i\boldsymbol{\beta}), \\ E[U_i\mathbf{b}_i\mathbf{b}_i^\top|\mathbf{Y}_i] &= \boldsymbol{\Lambda}_i + \boldsymbol{\Delta}_i(\bar{\mathbf{Z}}_i^2 + \boldsymbol{\gamma}_i\boldsymbol{\gamma}_i^\top\bar{\mathbf{Z}}_i^0 - \bar{\mathbf{Z}}_i^1\boldsymbol{\gamma}_i^\top - \boldsymbol{\gamma}_i\bar{\mathbf{Z}}_i^{1\top})\boldsymbol{\Delta}_i, \\ \mathbf{Z}_i|\mathbf{Y}_i &\sim t_{n_i}(\boldsymbol{\gamma}_i, \boldsymbol{\Omega}_i, \nu)\mathbb{I}_{\mathbb{B}_i}(\mathbf{Z}_i), \end{aligned} \tag{2.4.3}$$

em que $\bar{\mathbf{Z}}_i^0 = E\left[\frac{\nu+n_i}{\nu+\delta_i}|\mathbf{Y}_i\right]$, $\bar{\mathbf{Z}}_i^1 = E\left[\frac{\nu+n_i}{\nu+\delta_i}\mathbf{Z}_i|\mathbf{Y}_i\right]$, $\bar{\mathbf{Z}}_i^2 = E\left[\frac{\nu+n_i}{\nu+\delta_i}\mathbf{Z}_i\mathbf{Z}_i^\top|\mathbf{Y}_i\right]$, $\delta_i = (\mathbf{Z}_i - \boldsymbol{\gamma}_i)^\top\boldsymbol{\Omega}_i^{-1}(\mathbf{Z}_i - \boldsymbol{\gamma}_i)$, $\boldsymbol{\Delta}_i = \mathbf{D}\mathbf{W}_i^\top\boldsymbol{\Omega}_i^{-1}$, $\boldsymbol{\Lambda}_i = \mathbf{D} - \mathbf{D}\mathbf{W}_i^\top\boldsymbol{\Omega}_i^{-1}\mathbf{W}_i\mathbf{D}$, $\boldsymbol{\Omega}_i = \mathbf{W}_i\mathbf{D}\mathbf{W}_i^\top + \mathbf{I}_{n_i}$, $\boldsymbol{\gamma}_i = \mathbf{X}_i\boldsymbol{\beta}$, e $\mathbb{B}_i = B_{i1} \times \ldots, \times B_{in_i}$, em que B_{ij} pertence ao intervalo $(0, \infty)$ se $y_{ij} = 1$ ou pertence ao intervalo $(-\infty, 0]$ se $y_{ij} = 0$.

Por meio das equações (2.4.1)-(2.4.2), o passo E se reduz ao cálculo de $\bar{\mathbf{Z}}_i^0$, $\bar{\mathbf{Z}}_i^1$ e $\bar{\mathbf{Z}}_i^2$. Diretamente não se pode utilizar a Proposição 1 apresentada no Apêndice A com as expressões dadas em (2.4.3), já que as componentes associadas ao vetor aleatório $\mathbf{Z}_i|\mathbf{Y}_i$ são truncadas à direita ou à esquerda, dependendo dos valores de $y_{ij}, j = 1, \ldots, n_i$. Contudo, essas quantidades podem ser determinadas, em forma fechada, utilizando-se uma sequencia de transformações como as seguintes:

(i) O primeiro passo é a padronização das componentes de $\mathbb{B}_i$, como truncadas à direita ou à esquerda. Seja $\mathbf{A}_i$ uma matriz diagonal com elementos iguais a -1 ou 1 dependendo do intervalo $B_{ij} = (0, \infty)$ ou $B_{ij} = (-\infty, 0]$, respectivamente. Então, $\mathbf{U}_i \equiv \mathbf{A}_i\mathbf{Z}_i|\mathbf{Y}_i \sim Tt_{n_i}(\mathbf{A}_i\boldsymbol{\gamma}_i,\ \mathbf{A}_i\boldsymbol{\Omega}_i\mathbf{A}_i, \nu; \mathbb{C}_i)$, $\mathbb{C}_i = (-\infty, 0]^{n_i}$, ou seja, $\mathbf{U}_i$ segue uma distribuição multivariada t de Student $t_{n_i}(\mathbf{A}_i\boldsymbol{\gamma}_i, \mathbf{A}_i\boldsymbol{\Omega}_i\mathbf{A}_i, \nu)$ truncada à direita em $(-\infty, 0]^{n_i}$. Essa padronização facilita o cálculo de $\bar{\mathbf{U}}_i^0 = E\left[\frac{\nu + n_i}{\nu + \delta_i^u}|\mathbf{Y}_i\right]$, $\bar{\mathbf{U}}_i^1 = E\left[\frac{\nu + n_i}{\nu + \delta_i^u}\mathbf{U}_i|\mathbf{Y}_i\right]$, $\bar{\mathbf{U}}_i^2 = E\left[\frac{\nu + n_i}{\nu + \delta_i^u}\mathbf{U}_i\mathbf{U}_i^\top|\mathbf{Y}_i\right]$, por meio dos resultados apresentados na Proposição 1 do Apêndice A sobre as fórmulas dos primeiros dois momentos de uma distribuição multivariada t de Student, em que $\delta_i^u = (\mathbf{U}_i - \mathbf{A}_i\boldsymbol{\gamma}_i)^\top(\mathbf{A}_i\boldsymbol{\Omega}_i\mathbf{A}_i)^{-1}(\mathbf{U}_i - \mathbf{A}_i\boldsymbol{\gamma}_i)$.

(ii) Para o segundo passo basta notar que $\bar{\mathbf{Z}}_i^0 = \bar{\mathbf{U}}_i^0$, $\bar{\mathbf{Z}}_i^1 = \mathbf{A}_i^{-1}\bar{\mathbf{U}}_i^1$ e $\bar{\mathbf{Z}}_i^2 = \mathbf{A}_i^{-1}\bar{\mathbf{U}}_i^2\mathbf{A}_i^{-1}$, pois $\delta_i^u = \delta_i = (\mathbf{Z}_i - \boldsymbol{\gamma}_i)^\top\boldsymbol{\Omega}_i^{-1}(\mathbf{Z}_i - \boldsymbol{\gamma}_i)$.

Quando ν tende ao infinito, obtém-se um interessante algoritmo EM para o modelo probito como definido em (2.2.1).

2.4.1 Estimação da função de verossimilhança

A função de verossimilhança dos dados observados para $\boldsymbol{\theta} = (\boldsymbol{\beta}, \mathbf{D})$ é dada por (2.3.2)-(2.3.3). A integral associada a ela pode ser convenientemente calculada via amostragem por importância relacionada à qualquer distribuição contínua $\widetilde{\pi}$ (cujo suporte é maior que aquele associado a $\pi(\mathbf{b}_i, \boldsymbol{\theta}) \equiv t_q(\mathbf{b}_i|\mathbf{0}, \mathbf{D}, \nu)$). Então, a Equação (2.3.2) pode ser representada por

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{m} \int \psi_i^t(\boldsymbol{\beta}, \mathbf{b}_i)\frac{\pi(\mathbf{b}_i, \boldsymbol{\theta})}{\widetilde{\pi}(\mathbf{b}_i, \boldsymbol{\theta})}\widetilde{\pi}(\mathbf{b}_i, \boldsymbol{\theta})d\mathbf{b}_i,$$

em que $\psi_i^t(\boldsymbol{\beta}, \mathbf{b}_i)$ foi definido em (2.3.3). Consequentemente, $\ell(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta})$ pode ser estimada sem nenhum custo computacional por

$$\widehat{\ell(\boldsymbol{\theta})} \doteq \sum_{i=1}^{m} \log\left[\frac{1}{K}\sum_{l=1}^{K}[\psi_i(\boldsymbol{\beta}, \mathbf{b}_i^{(l)})\frac{\pi(\mathbf{b}_i^{(l)}, \boldsymbol{\theta})}{\widetilde{\pi}(\mathbf{b}_i^{(l)}, \boldsymbol{\theta})}]\right].$$

em que $\mathbf{b}_1, \ldots, \mathbf{b}_l, \ldots, \mathbf{b}_K$ são amostradas de $\widetilde{\pi}(\mathbf{b}_i, \boldsymbol{\theta})$. Uma escolha eficiente para $\widetilde{\pi}$ consiste na distribuição condicional de $\mathbf{b}_i$ dadas as observações $\mathbf{y}_i$, $i = 1, \ldots, m$ (Robert et al., 1999).

Outra abordagem para se calcular a verossimilhança é a de se considerar a função *adaptIntegrate* do pacote *cubature* (Johnson and Narasimhan, 2011) disponível no *software* R. Essa função é apropriada para se calcular integrais multivariadas numericamente e foi utilizada para se obter os valores da função log-verossimilhança aplicados aos estudos empíricos apresentados neste trabalho.

A função log-verossimilhança pode ser utilizada para se monitorar a convergência do algoritmo EM. Na prática, as iterações do algoritmo EM devem ser repetidas até que o valor associado a alguma função de distância envolvendo duas sucessivas avaliações da função log-verossimilhança $\ell(\boldsymbol{\theta})$, tais como $||\ell(\widehat{\boldsymbol{\theta}}^{(k+1)}) - \ell(\widehat{\boldsymbol{\theta}}^{(k)})||$ ou $||\ell(\widehat{\boldsymbol{\theta}}^{(k+1)})/\ell(\widehat{\boldsymbol{\theta}}^{(k)}) - 1||$, $k = 0, 2, ...$, seja suficientemente pequeno. Adicionalmente, os critérios de seleção de modelos baseados na função de verossimilhança observada, tais como o AIC e BIC, podem ser obtidos.

2.4.2 Estimação dos erros-padrão

Denote as estimativas MV do algoritmo EM por $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\beta}}, \widehat{\mathbf{D}})$. Sob certas condições gerais de regularidade, pode-se obter um método baseado na função de informação, seguindo McLachlan and Krishnan (1997), para se computar a matriz de covariância assintótica das estimativas do modelo t-GLMM para dados correlacionados. Defina $I_o(\widehat{\boldsymbol{\theta}}|\mathbf{Y}) = \sum_{i=1}^{m} \widehat{\mathbf{u}}_i\widehat{\mathbf{u}}_i^\top$ como sendo a matriz de informação observada, em que $\widehat{\mathbf{u}}_i = E\left[\frac{\partial \ell_i(\boldsymbol{\theta}|\mathbf{Y}_{comp})}{\partial \boldsymbol{\theta}}|\mathbf{y}_i\right]\Big|_{\widehat{\boldsymbol{\theta}}} = \frac{\partial Q_i(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}}\Big|_{\widehat{\boldsymbol{\theta}}}$, com a função de verossimilhança dos dados completos $\ell_i(\boldsymbol{\theta}|\mathbf{Y}_{comp})$ avaliada para uma única observação $\mathbf{y}_i$, $i = 1, \ldots, m$.

Expressões explícitas para os elementos de $\widehat{\mathbf{u}}_i$ são dadas por:

$$
\begin{aligned}
\frac{\partial Q_i(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})}{\partial \boldsymbol{\beta}}\Big|_{\widehat{\boldsymbol{\theta}}} &= \mathbf{X}_i^\top(\widehat{U_i\mathbf{Z}_i} - \mathbf{W}_i\widehat{U_i\mathbf{b}_i}) - (\mathbf{X}_i\mathbf{X}_i^\top)\widehat{U}_i\widehat{\boldsymbol{\beta}} \\
\frac{\partial Q_i(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})}{\partial d_r}\Big|_{\widehat{\boldsymbol{\theta}}} &= -\frac{1}{2}\left[tr\left(\widehat{\mathbf{D}}^{-1}\frac{\partial \widehat{\mathbf{D}}}{\partial d_r}\widehat{\mathbf{D}}^{-1}(\widehat{\mathbf{D}} - \widehat{\mathbf{b}_i\mathbf{b}_iU_i})\right)\right]
\end{aligned}
$$

em que $\widehat{U}_i = E[U_i|\mathbf{Y}_i]\big|_{\widehat{\boldsymbol{\theta}}} = \bar{\mathbf{Z}}_i^0\big|_{\widehat{\boldsymbol{\theta}}}$, $\widehat{U_i\mathbf{Z}_i} = E[U_i\mathbf{Z}_i|\mathbf{Y}_i]\big|_{\widehat{\boldsymbol{\theta}}} = \bar{\mathbf{Z}}_i^1\big|_{\widehat{\boldsymbol{\theta}}}$, $\widehat{U_i\mathbf{b}_i} = E[U_i\mathbf{b}_i|\mathbf{Y}_i]\big|_{\widehat{\boldsymbol{\theta}}}$, $\widehat{\mathbf{b}_i\mathbf{b}_iU_i} = E[\mathbf{b}_i\mathbf{b}_i^\top U_i|\mathbf{Y}_i]\big|_{\widehat{\boldsymbol{\theta}}}$, d_r são os elementos distintos da matriz quadrada $\mathbf{D}$. Os erros-padrão são obtidos pela raiz quadrada dos elementos diagonais da inversa da matriz de informação estimada $I_o(\widehat{\boldsymbol{\theta}}|\mathbf{Y})$.

2.4.3 Estimação dos efeitos aleatórios

Nesta seção, consideramos uma formulação via estimação Bayesiana empírica para os efeitos aleatórios. Da distribuição condicional de $\mathbf{b}_i$ dada $(\mathbf{Y}_i, \mathbf{Z}_i, U_i)$ (veja o Apêndice A), tem-se:

$$
f(\mathbf{b}_i|\mathbf{Y}_i, \mathbf{Z}_i, u_i, \boldsymbol{\theta}) = f(\mathbf{b}_i|\mathbf{Z}_i, u_i, \boldsymbol{\theta}) = \phi_q(\mathbf{b}_i|\boldsymbol{\Delta}_i(\mathbf{Z}_i - \mathbf{X}_i^\top\boldsymbol{\beta}), u_i^{-1}\boldsymbol{\Lambda}_i),
$$

em que $\boldsymbol{\Delta}_i = \mathbf{D}\mathbf{W}_i^\top\boldsymbol{\Omega}_i^{-1}$, $\boldsymbol{\Lambda}_i = \mathbf{D} - \mathbf{D}\mathbf{W}_i^\top\boldsymbol{\Omega}_i^{-1}\mathbf{W}_i\mathbf{D}$ e $\boldsymbol{\Omega}_i = \mathbf{W}_i\mathbf{D}\mathbf{W}_i^\top + \mathbf{I}_{n_i}$, $i = 1, \ldots, m$.

Dessa forma, o menor erro quadrático médio (MSE) do preditor de $\mathbf{b}_i$, obtido pela média condicional de $\mathbf{b}_i$ dado $\mathbf{Y}_i = \mathbf{y}_i$, é

$$
\hat{\mathbf{b}}_i = E\left[\mathbf{b}_i|\mathbf{Y}_i, \boldsymbol{\theta}\right] = E\left[E(\mathbf{b}_i|\mathbf{Y}_i, \mathbf{Z}_i, \boldsymbol{\theta})|\mathbf{Y}_i, \boldsymbol{\theta}\right] = \boldsymbol{\Delta}_i[E(\mathbf{Z}_i|\mathbf{Y}_i, \boldsymbol{\theta}) - \mathbf{X}_i^\top\boldsymbol{\beta}],
$$

em que $E[\mathbf{Z}_i|\mathbf{Y}_i, \boldsymbol{\theta}]$ é o valor esperado de uma distribuição multivariada truncada t de Student apresentada em (2.4.3).

2.5 Aplicações

Nesta seção será apresentado um resumo da aplicação dos métodos desenvolvidos. Para tanto, um estudo com dados simulados e um banco de dados reais referente a um estudo de coorte de doenças respiratórias em 111 pacientes serão utilizados.

Estudo de simulação

O objetivo principal do estudo de dados simulados foi o de avaliar o desempenho dos métodos propostos, investigando os efeitos na inferência dos parâmetros quando a suposição de normalidade é violada na função de ligação do modelo, bem como para os efeitos aleatórios. Adicionalmente, esse estudo tem como propósito de analisar se os critérios de comparação de modelos (AIC e BIC) conseguem selecionar corretamente o modelos para os dados simulados.

Para tanto, definiu-se um estudo balanceado com seis variáveis respostas, $j = 1, \ldots, 6$ e o seguinte modelo GLMM para dados binários:

$$
\begin{aligned}
Y_{ij} = \mathbb{I}_{(Z_{ij}>0)},\ \mathbf{Z}_i|\mathbf{b}_i, U_i = u_i &\sim N_{ni}(\boldsymbol{\mu}_i, u_i^{-1}\mathbf{I}_{n_i}),\\
\mathbf{b}_i|U_i = u_i &\sim N_q(\mathbf{0}, u_i^{-1}\mathbf{D}),\\
U_i &\sim F(\nu),
\end{aligned}
\tag{2.5.1}
$$

em que $\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{i6})^\top$, com $\mu_{ij} = \beta_0 + \beta_1 X_{i1} + b_{1i} + b_{2i}X_{i2}$ para todo j. Uma covariável binária X_1 foi gerada para os efeitos fixos e uma covariável normal padrão X_2, para os efeitos aleatórios; $F(\nu)$ é uma função de distribuição positiva. Finalmente, foram fixados $\beta_0 = 1$, $\beta_1 = 0.8$ e uma matriz $\mathbf{D}$ dada por $\mathbf{D} = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$ para representar a estrutura de correlação entre as observações associadas ao i-ésimo indivíduo.

Neste estudo, três tamanhos de amostras foram gerados (n=100, 250 e 500), 1000 réplicas de cada estudo foram avaliadas e ajustou-se, concomitantemente, os modelos probito-GLMM (Subseção 2.2) e o t-GLMM com 4 graus de liberdade (Subseção 2.3). Esse estudo foi divido em três cenários e, a seguir, um resumo dos resultados obtidos será feito.

Cenário 1: Nesse cenário, os dados foram simulados da função de ligação probito (modelo probito-GLMM) e assumiu-se também a normalidade para os efeitos aleatórios. Ou seja, no modelo (2.5.1) $U = 1$ seguiu uma distribuição degenerada com $P(U = 1) = 1$. Pela análise das estimativas dos efeitos fixos (β_0 e β_1) (Tabela 2.1), verificou-se que os modelos probito-GLMM e t-GLMM conseguiram recuperar bem os valores verdadeiros desses parâmetros. As estimativas das medidas de erro-padrão foram adequadas, uma vez que a abordagem via matriz de informação observada (Subseção 2.4.2) foi próxima às estimativas encontradas via Monte Carlo. Quanto maior foi o tamanho da amostra, melhor foram as estimativas dos parâmetros segundo os dois modelos avaliados, bem como menores medidas de erro-padrão, como era esperado. Segundo os critérios de comparação de modelos (AIC e BIC), o modelo probito-GLMM foi corretamente escolhido.

Cenário 2: Para esse cenário, os dados foram simulados da função de ligação t de Student e assumiu-se a distribuição t de Student para os efeitos aleatórios com $\nu = 4$. Ou seja, em (2.5.1), tem-se: $U \sim \text{Gamma}(\nu/2, \nu/2)$. Observando as estimativas dos modelos probito-GLMM e t-GLMM para esses dados (Tabela 2.2), verificou-se que as menores diferenças entre os valores estimados e os verdadeiros foram

Tabela 2.1: **Cenário 1.** MC mean, MC Sd (em parênteses) e MC Coverage são, respectivamente, as estimativas médias via Monte Carlo, os desvios-padrão e as medidas médias da proporção de cobertura ao se ajustar os modelos probito-GLMM e t-GLMM. IM SE são os valores médios das estimativas de erro-padrão aproximados obtidos pelo método baseado na função de informação. MC AIC e MC BIC são os valores médios das respectivos critérios de comparação de modelos.

			Dados simulados probito-GLMM			
n	Modelo		β_0	β_1	MC AIC	MC BIC
100	probito-GLMM	MC Mean	1.006	0.843	554.59	570.22
		IM SE	0.202	0.478		
		MC Sd	(0.171)	(0.399)		
		MC Coverage	98%	98%		
	t-GLMM	MC Mean	1.062	0.933	556.23	571.89
		IM SE	0.217	0.612		
		MC Sd	(0.179)	(0.428)		
		MC Coverage	99%	99%		
250	probito-GLMM	MC Mean	1.007	0.815	1371.69	1392.81
		IM SE	0.127	0.280		
		MC Sd	(0.125)	(0.246)		
		MC Coverage	97%	97%		
	t-GLMM	MC Mean	1.061	0.899	1375.99	1397.12
		IM SE	0.136	0.363		
		MC Sd	(0.133)	(0.272)		
		MC Coverage	96%	99%		
500	probito-GLMM	MC Mean	1.005	0.820	2734.52	2742.98
		IM Sd	0.090	0.195		
		MC SE	(0.106)	(0.172)		
		MC Coverage	91%	97%		
	t-GLMM	MC Mean	1.060	0.902	2759.80	2768.27
		IM SE	0.096	0.253		
		MC Sd	(0.114)	(0.196)		
		MC Coverage	88%	98%		

Tabela 2.2: **Cenário 2.** MC mean, MC Sd (em parênteses) e MC Coverage são, respectivamente, as estimativas médias via Monte Carlo, os desvios-padrão e as medidas médias da proporção de cobertura ao se ajustar os modelos probito-GLMM e t-GLMM. IM SE são os valores médios das estimativas de erro-padrão aproximados obtidos pelo método baseado na função de informação. MC AIC e MC BIC são os valores médios das respectivos critérios de comparação de modelos.

			Dados simulados t-GLMM			
n	Modelo		β_0	β_1	MC AIC	MC BIC
100	probito-GLMM	MC Mean	0.952	0.740	576.46	592.10
		IM SE	0.207	0.467		
		MC Sd	(0.183)	(0.423)		
		MC Coverage	96%	97%		
	t-GLMM	MC Mean	1.004	0.831	574.23	589.86
		IM SE	0.221	0.580		
		MC Sd	(0.192)	(0.452)		
		MC Coverage	98%	98%		
250	probito-GLMM	MC Mean	0.955	0.723	1426.14	1447.27
		IM SE	0.131	0.276		
		MC Sd	(0.119)	(0.239)		
		MC Coverage	94%	96%		
	t-GLMM	MC Mean	1.004	0.813	1420.96	1442.09
		IM SE	0.140	0.345		
		MC Sd	(0.126)	(0.261)		
		MC Coverage	97%	99%		
500	probito-GLMM	MC Mean	0.956	0.717	2846.13	2871.42
		IM SE	0.092	0.191		
		MC Sd	(0.106)	(0.177)		
		MC Coverage	88%	95%		
	t-GLMM	MC Mean	1.004	0.809	2836.01	2861.30
		IM SE	0.098	0.240		
		MC Sd	(0.117)	(0.200)		
		MC Coverage	91%	99%		

encontradas quando o ajuste foi feito pelo modelo t-GLMM. Como no cenário anterior, AIC e BIC corretamente selecionam o modelo em que os dados foram gerados.

Cenário 3: Aqui, os dados foram simulados da função de ligação Normal Contaminada e os efeitos aleatórios seguiram uma distribuição Normal Contaminada com $\nu = (\nu_1, \nu_2)^\top = (0.1, 0.1)^\top$. Ou seja, em (2.5.1), $U = \begin{cases} \nu_2 & \text{com prob} \quad \nu_1 \\ 1 & \text{com prob} \quad 1-\nu_1 \end{cases}$. Com base na Tabela 2.3, o modelo que apresentou melhor performance foi o t-GLMM. Menores valores de AIC e BIC, além de melhores estimativas, foram obtidos para o modelo t-GLMM e esses resultados reafirmam que esta função de ligação é mais robusta a desvios da suposição de normalidade do modelo que a probito e também fornecem fortes indícios de que o algoritmo proposto é eficiente na estimação dos parâmetros do modelo.

Tabela 2.3: **Cenário 3.** MC mean, MC Sd (em parênteses) e MC Coverage são, respectivamente, as estimativas médias via Monte Carlo, os desvios-padrão e as medidas médias da proporção de cobertura ao se ajustar os modelos probito-GLMM e t-GLMM. IM SE são os valores médios das estimativas de erro-padrão aproximados obtidos pelo método baseado na função de informação. MC AIC e MC BIC são os valores médios das respectivos critérios de comparação de modelos.

			Dados simulados da Normal Contaminada			
n	Modelo		β_0	β_1	MC AIC	MC BIC
100	probito-GLMM	MC Mean	0.880	0.699	592.13	607.77
		IM SE	0.192	0.430		
		MC Sd	(0.164)	(0.365)		
		MC Coverage	92%	96%		
	t-GLMM	MC Mean	0.931	0.785	589.86	605.49
		IM SE	0.206	0.535		
		MC Sd	(0.172)	(0.393)		
		MC Coverage	96%	98%		
250	probito-GLMM	MC Mean	0.886	0.683	1463.82	1484.95
		IM SE	0.121	0.257		
		MC Sd	(0.116)	(0.241)		
		MC Coverage	84%	93%		
	t-GLMM	MC Mean	0.936	0.774	1458.92	1480.05
		IM SE	0.130	0.323		
		MC Sd	(0.124)	(0.264)		
		MC Coverage	93%	97%		
500	probito-GLMM	MC Mean	0.879	0.687	2921.92	2947.20
		IM SE	0.085	0.179		
		MC Sd	(0.095)	(0.171)		
		MC Coverage	68%	90%		
	t-GLMM	MC Mean	0.931	0.780	2911.85	2937.13
		IM SE	0.092	0.226		
		MC Sd	(0.105)	(0.197)		
		MC Coverage	84%	97%		

Tabela 2.4: Dados reais. Estimativas MV e erros-padrão para os dados reais segundo os modelos com função de ligação probito e t-GLMM.

		probito-GLMM		t-GLMM	
Variável	Parâmetro	MV	EP	MV	EP
intercept	β_0	-1.0992	0.5151	-1.1126	0.5965
treat	β_1	-0.9153	0.2524	-1.0589	0.2944
gender	β_2	0.1278	0.2970	0.0873	0.3447
center	β_3	0.5700	0.2415	0.6175	0.2751
baseline	β_4	1.6202	0.2650	1.7613	0.3061
covariance	d_{11}	0.0163	0.0061	0.0134	0.0053
AIC		575.5527		562.9703	
BIC		606.9342		594.3517	

Aplicação a dados reais

Esta base foi apresentada por Halekoh et al. (2006) e se refere a um estudo de coorte sobre uma doença respiratória em 111 pacientes de dois centros clínicos em 4 visitas. Para cada visita, o estado respiratório de cada paciente era classificado como bom (= 1) ou ruim (= 0). Os pacientes foram selecionados aleatoriamente para receber tratamento ativo ou o placebo. Para tanto, assume-se que a probabilidade $Pr(Y_{ij} = 1|b_i)$ de infecção respiratória do i-ésimo paciente na j-ésima visita, $i = 1, \ldots, 111$, $j = 1, \ldots, 4$, é dada por:

$$F(\beta_0 + \text{treat}\,\beta_1 + \text{gender}\,\beta_2 + \text{center}\,\beta_3 + \text{baseline}\,\beta_4 + age\,b_i),$$

em que "F" é a função de ligação probito-GLMM ou t-GLMM; "treat" é o tratamento do paciente, 1 se o tratamento for ativo e 0 se for placebo; "gender = 0" para sexo feminino e 1 para masculino; "center = 0" para o primeiro centro e 1 para o segundo; e "baseline" é a resposta na primeira visita. Os elementos da matriz $\mathbf{W}$ são w_{ij} = age, em que, a idade do paciente incluído no estudo foi centralizada em torno de 31 anos (valor mediano). Essa definição para os efeitos aleatórios indica que o efeito da idade está intrinsecamente associado ao paciente (d_{11}).

Uma avaliação prévia do impacto dos graus de liberdade na função log-verossimilhança foi feita para se obter a melhor estimativa para esse parâmetro. Por meio dessa análise, verificou-se que o valor de ν que melhor maximiza a verossimilhança era 4 e, portanto, esse valor foi fixado para as análises subsequentes.

Por meio da Tabela 2.4, verificou-se que as estimativas via máxima verossimilhança para ambos os modelos avaliados foram próximas. Contudo, o que obteve melhor desempenho segundo os critério AIC e BIC foi o t-GLMM. Segundo esse modelo, a probabilidade de se ter a doença respiratória está associada ao tipo de tratamento, tipo de centro médico e a resposta do paciente na primeira visita (baseline). Com relação à convergência do algoritmo, observou-se que ambas as abordagens atingiram a convergência e, para todos as iterações do algoritmo EM proposto, a performance do modelo t-GLMM foi melhor que a do modelo probito.

Capítulo3

Modelos de Análise de Fatorial Confirmatória Tobit

Neste capítulo apresentaremos um resumo dos resultados de dois artigos relacionados ao modelo de Análise de Fatorial Confirmatória Tobit. O primeiro, intitulado "Estimation Methods for Multivariate Tobit Confirmatory Factor Analysis", foi submetido à apreciação da revista *Computational Statistics and Data Analysis* e está descrito no Apêndice B. Já o segundo, denominado "Likelihood-Based Inference for Tobit Confirmatory Factor Analysis Using the Multivariate t-Distribution", apresentado no Apêndice C, foi submetido para publicação no *Statistics and Computing*. Vale ainda destacar que o primeiro trabalho contou com a colaboração dos professores Dr. Jorge L. Bazán da Universidade de São Paulo/São Carlos e Dr. Caio L. N. Azevedo da Unicamp.

3.1 Introdução

A Análise Fatorial (AF) é um dos métodos mais populares de análise estatística multivariada. Ela tem por objetivo fazer inferências sobre um conjunto de variáveis em termo de um número menor de variáveis latentes, chamadas de fatores. Os primeiros trabalhos de AF são dados a Spearman (1904), mas foram Lawley and Maxwell (1971) que definiram uma abordagem mais teórica e estatística para a área. A partir de então, diversos trabalhos têm sido publicados sobre o tema, Mulaik (1972) e Bartholomew et al. (2011) são alguns exemplos. Nessa área é comum a definição de modelos de análise fatorial exploratória (EFA) e modelos de análise fatorial confirmatória (CFA). Enquanto o primeiro procura identificar a quantidade ideal de fatores latentes, avaliar a relação entre as variáveis observadas e os fatores, nos modelos CFA essas definições são feitas previamente e a análise preocupa-se essencialmente em confirmar a estrutura proposta (Jöreskog, 1969). Neste trabalho o foco será nos modelos CFA.

Inicialmente, os modelos de AF foram desenvolvidos para dados contínuos e normalmente distribuídos. Todavia, em algumas situações faz-se necessária a análise de dados censurados. Ignorando a informação associada aos casos censurados, a análise produzirá uma representação viesada dos dados, uma vez que o mecanismo que gera essess dados pode conter informação sobre a estrutura do fator (Kamakura and Wedel, 2001). Adicionalmente, os resultados obtidos pela análise fatorial clássica serão viesados, já que a suposição de normalidade não será mais atendida (Muthén, 1989). Desta forma, modelos Tobit têm recebido recentemente interesse nessa área.

O modelo Tobit desenvolvido por Tobin (1958) para dados censurados em zero tem a vantagem de oferecer uma ligação explícita entre o mecanismo de geração dos dados censurados e não censurados por permitir uma gama de especificações sobre as variáveis latentes e restringindo a distribuição dos dados não censurados a uma distribuição com suporte positivo. Por causa da sua flexibilidade na modelagem de dados mistos, os modelos Tobit têm recebido recentemente muita atenção na literatura estatística. O primeiro a introduzir os modelos Tobit na análise fatorial foi Muthén (1989). Mais tarde, Waller and Muthén (1992), Huang (1999), Kamakura and Wedel (2001), Zhou and Liu (2009) desenvolveram diferentes procedimentos de análise de dados multivariados através da Análise Fatorial Confirmatória Tobit (TCFA). Diferentemente da abordagem desses autores, neste capítulo apresentamos métodos de estimação clássica do modelo TCFA sob distribuição Normal e t de Student, além de apresentar uma modelagem via análise Bayesiana para o modelo TCFA Gaussiano. Desta forma, este capítulo está dividido basicamente em dois blocos: o primeiro que retrata o desenvolvimento de métodos (frequentistas e Bayesianos) para o modelo TCFA sob normalidade e o segundo, que propõe um método frequentista para o modelo TCFA supondo distribuição t de Student para os dados.

A principal motivação para o estudo do modelo TCFA sob distribuição Normal se baseia na proposta de se apresentar métodos alternativos de análise, além de fornecer instrumentos adicionais para análise de diagnóstico de dados influentes nesta abordagem. Como no capítulo 2, a primeira abordagem a ser apresentada é o desenvolvimento de um algoritmo ECM exato, em que ambos os passos E e M são obtidos de forma direta. Diferentemente dos trabalhos de Huang (1999) e de Zhou and Liu (2009) que implementaram um algoritmo MCEM, aqui derivamos fórmulas fechadas para as expressões do passo E que se reduzem ao cálculo dos primeiros dois momentos de uma distribuição multivariada Normal truncadas que podem ser obtidos utilizando pacotes específicos do *software* `R`. Por meio dessa perspectiva, faz-se possível avaliar numericamente a função de verossimilhança, o que possibilita sua utilização no monitoramento da convergência do algoritmo ou no auxílio do cálculo de critérios de comparação de modelos, tais como AIC, BIC e CAIc (Bozdogan, 1987).

A segunda abordagem do estudo do modelo Gaussiano TCFA está relacionada ao desenvolvimento de procedimento completamente Bayesiano. Recentes desenvolvimentos de métodos de Monte Carlo Cadeia de Markov (MCMC) permitem uma implementação fácil e direta de uma análise Bayesiana por meio de *softwares*, tais como o `OpenBugs`, desde que uma estrutura hierárquica do modelo esteja disponível. Esta persepectiva permite uma grande flexibilidade para ajuste de modelos a dados de diferentes graus de complexidade (Dunson, 2001), fazendo o uso de toda informação disponível no estudo, acomodando na estimação paramétrica a incerteza associados aos parâmetros por meio de uso de distribuições *a priori* apropriadas, fornecendo afirmações probabilísticas diretas a respeito dos parâmetros por meio de intervalos de crebilidade, não dependendo assim de resultados assintóticos. Adicionalmente, propomos medidas de diagnóstico de dados influentes para o modelo TCFA por esta abordagem. Esta análise permite avaliar o impacto de certas observações amostrais na estimação paramétrica.

Por entender que modelos Gaussianos são sensíveis a pontos discrepantes, neste trabalho também propomos uma abordagem robusta utilizando uma distribuição de caudas mais pesadas que a Normal, a distribuição t de Student. Zhang et al. (2013) propuseram a utilização dessa distribuição no modelo de análise fatorial clássica, denominado aqui por modelo t-CFA. Na presença de dados censurados, por outro lado, Zhou and Liu (2009) desenvolveram um algoritmo Monte Carlo EM (MCEM) para o modelo TCFA sob a suposição de que a variável latente segue uma distribuição multivariada t de Student. Diferentemente desse autores, propomos neste trabalho um algoritmo ECM exato por meio da adaptação da estrutura do algoritmo desenvolvido para o modelo TCFA Gaussiano.

Este capítulo está organizado da seguinte forma: na Seção 3.2, descrevemos o modelo de Análise Fatorial Confirmatório Tobit Gaussiano, além disso, apresentamos os métodos de estimação via máxima verossimilhança e via análise Bayesiana. Ainda nessa seção apresentamos métodos de diagnóstico sob perspectiva Bayesiana para análise de dados influentes. Os métodos propostos foram avaliados empiricamente por meio de um estudo simulado e uma aplicação a dados reais que, resumidamente, são descritos na Subseção 3.2.3. Já o modelo de Análise Fatorial Confirmatório Tobit com distribuição t (t-TCFA), bem como os métodos inferenciais propostos para esse modelo e sua aplicação, serão apresentados na Seção 3.3

3.2 Modelo Gaussiano TCFA

Seja $\mathbf{y}_i = (y_{i1}, \ldots, y_{in_i})^\top$ um vetor $(p \times 1)$ de respostas contínuas para o sujeito i, $i = 1, ..., n$ em que cada resposta y_{ij} toma valores não-negativos. O modelo de Análise Fatorial Gaussiano com covariáveis (CFA) é dado por (Zhou and Liu, 2009):

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\Lambda}\mathbf{z}_i + \boldsymbol{\epsilon}_i, \tag{3.2.1}$$

em que $\mathbf{z}_i \overset{iid}{\sim} N_q(\mathbf{0}, \boldsymbol{\Omega})$ é independente de $\boldsymbol{\epsilon}_i \overset{ind.}{\sim} N_p(\mathbf{0}, \boldsymbol{\Psi})$, $\quad i = 1, \ldots, m$; o subscrito i refere-se ao indivíduo; $\mathbf{y}_i = (y_{i1}, \ldots, y_{ip})^\top$ é um vetor de dimensão $(p \times 1)$ de respostas contínuas para o indivíduo i; $\mathbf{X}_i$ é a matriz de delineamento de dimensão $(p \times k)$ que possui a seguinte caracterização:

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

com $\mathbf{x}_{ij}$ sendo um vetor de dimensão $(k \times 1)$; $\boldsymbol{\beta}$, são os efeitos fixos de dimensão $(k \times 1)$; $\boldsymbol{\Lambda}$ é a matriz de cargas fatoriais de dimensão $(p \times q)$; $\mathbf{z}_i$ é um vetor $(q \times 1)$ $(q < p)$ dos fatores aleatórios ; $\boldsymbol{\epsilon}_i$ é um vetor $(p \times 1)$ dos erros aleatórios, $\boldsymbol{\Omega}$ é a matriz de covariância/correlação dos fatores, geralmente fixada previamente na análise confirmatória e $\boldsymbol{\Psi}$ uma matriz diagnoal de covariância dos erros.

Na presente formulação, considera-se ainda que a resposta Y_{ij} não é completamente observada para todo i, j. Seguindo Vaida and Liu (2009) e Matos et al. (2013a), considere os dados observados para o i-ésimo indivíduo por $(\mathbf{V}_i, \mathbf{C}_i)$, em que $\mathbf{V}_i$ representa o vetor de casos não censurados (ou seja, quando a observação y_{ij} não é censurada, V_{ij} será exatamente ao valor observado) ou será o nível de censura (quando a observação y_{ij} é censurada, V_{ij} será uma constante associada ao nível de censura) e $\mathbf{C}_i$ é o vetor indicador de censura com componentes C_{ij}. Dessa forma, pode-se resumir os dados observados da seguinte forma:

$$
\begin{aligned}
y_{ij} &\leq V_{ij} \quad \text{se} \quad C_{ij} = 1, \\
y_{ij} &= V_{ij} \quad \text{se} \quad C_{ij} = 0.
\end{aligned} \tag{3.2.2}
$$

Com a combinação de (3.2.1) e (3.2.2), o modelo Gaussiano de Análise de Fatorial Tobit com covariáveis (TCFA) está caracterizado. Essa formulação do modelo com (3.2.2) é uma generalização da proposta de Zhou and Liu (2009), uma vez que em nosso caso o nível de censura V_{ij} pode assumir qualquer valor e o nível de censura igual a zero representa um caso particular. Cabe ainda notar que o tipo de censura aqui definido refere-se à censura à esquerda. Extensão para um nível de censura à direita ou arbitrária é imediata.

Em geral, modelos de análise fatorial não são identificados, já que se for considerado $\boldsymbol{\Lambda}^* = \boldsymbol{\Lambda}\boldsymbol{\Gamma}^{-1}$ e $\mathbf{z}_i^* = \boldsymbol{\Gamma}\mathbf{z}_i$, para qualquer matriz não singular $\boldsymbol{\Gamma}$, será obtido o mesmo modelo apresentado na equação (3.2.1). Para resolução desse problema, Zhou and Liu (2009), por exemplo, indicam a fixação de alguns elemententos da matriz $\boldsymbol{\Lambda}$ e/ou $\boldsymbol{\Omega}$. Dada a natureza confirmatória da análise, a matriz $\boldsymbol{\Omega}$ será considerada conhecida em todo desenvolvimento do trabalho, a fim de se evitar a não identificabilidade do modelo.

3.2.1 Estimação por Máxima Verossimilhança

Dado o vetor de parâmetros $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \boldsymbol{\Lambda}, \Psi)^\top$, métodos de inferência clássica se baseiam na distribuição marginal de $\mathbf{y}_i$, $i = 1, \ldots, m$. Quando os dados não são censurados, $\mathbf{y}_i \overset{\text{ind.}}{\sim} N_p(\mathbf{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma})$, em que $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Omega}\boldsymbol{\Lambda}^\top + \boldsymbol{\Psi}$. Todavia, para respostas censuradas, tal como definido em (3.2.2), tem-se que $\mathbf{y}_i \sim TN_p(\mathbf{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma}_i; \mathbb{A})$, em que $TN_p(.; \mathbb{A})$ denota uma distribuição Normal truncada no intervalo $\mathbb{A}$, em que $\mathbb{A}_i = A_{i1} \times \ldots, \times A_{in_i}$, com A_{ij} pertencente ao intervalo $(-\infty, \infty)$, se $C_{ij} = 0$ ou pertencente a $(-\infty, 0]$, se $C_{ij} = 1$.

Para se obter a função de verossimilhança associada ao modelo TCFA, primeiramente, serão tratadas separadamente as componentes censuradas e observadas do vetor $\mathbf{y}_i$. Seguindo Vaida and Liu (2009), sejam $\mathbf{y}_i^o$ o vetor das respostas observadas de dimensão p^o e $\mathbf{y}_i^c$ o vetor das observações censuradas p^c-dimensional para o indivíduo i em que $(p = p^o + p^c)$, tal que, $C_{ij} = 0$ para todos os elementos em $\mathbf{y}_i^o$ e 1 para todos elementos em $\mathbf{y}_i^c$. Após esta ordenação, $\mathbf{y}_i$, $\mathbf{V}_i$, $\mathbf{X}_i$ e $\boldsymbol{\Sigma}_i$ podem ser particionadas como

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

em que $vec(.)$ denota a transformação linear que converte uma matriz em um vetor coluna. Desta forma, tem-se que $\mathbf{y}_i^o \sim N_{p^o}(\mathbf{X}_i^o\boldsymbol{\beta}, \boldsymbol{\Sigma}_i^{oo})$, $\mathbf{y}_i^c|\mathbf{y}_i^o \sim N_{p^c}(u_i, \mathbf{S}_i)$, em que $\boldsymbol{u_i} = \mathbf{X}_i^c\boldsymbol{\beta} + \boldsymbol{\Sigma}_i^{co}(\boldsymbol{\Sigma}_i^{oo})^{-1}(\mathbf{y}_i^o - \mathbf{X}_i^o\boldsymbol{\beta})$ e $\mathbf{S}_i = \boldsymbol{\Sigma}_i^{cc} - \boldsymbol{\Sigma}_i^{co}(\boldsymbol{\Sigma}_i^{oo})^{-1}\boldsymbol{\Sigma}_i^{oc}$. Agora, seja $\Phi_p(\mathbf{u}; \mathbf{a}, \mathbf{A})$ e $\phi_p(\mathbf{u}; \mathbf{a}, \mathbf{A})$ a função de distribuição acumulada e a função de densidade, respectivamente, de $N_p(\mathbf{a}, \mathbf{A})$ avaliada em $\mathbf{u}$. De Vaida and Liu (2009) e Matos et al. (2013a), a função de verossimilhança associada à observação i (utilizando as propriedades de probabilidade condicional) é dada por

$$
\begin{aligned}
L_i(\boldsymbol{\theta}) = f(\mathbf{y}_i|\boldsymbol{\theta}) &= P(\mathbf{V}_i|\mathbf{C}_i, \boldsymbol{\theta}) = P(\mathbf{y}_i^c \leq \mathbf{V}_i^c|\mathbf{y}_i^o = \mathbf{V}_i^o, \boldsymbol{\theta})P(\mathbf{y}_i^o = \mathbf{V}_i^o|\boldsymbol{\theta}), \\
&= P(\mathbf{y}_i^c \leq \mathbf{V}_i^c|\mathbf{y}_i^o, \boldsymbol{\theta})f(\mathbf{y}_i^o|\boldsymbol{\theta}) \\
&= \phi_{p^o}(\mathbf{y}_i^o; \mathbf{X}_i^o\boldsymbol{\beta}, \boldsymbol{\Sigma}_i^{oo})\Phi_{p^c}(\mathbf{V}_i^c; u_i, \mathbf{S}_i) = \alpha_i,
\end{aligned} \tag{3.2.3}
$$

que pode ser obtida sem muito esforço computacional, por meio, por exemplo, do pacote `mvtnorm` disponível no software R (veja, por exemplo, Genz et al., 2008; R Core Team, 2012). A função log-verossimilhança para os dados observados é, então, dada por $\ell(\boldsymbol{\theta}|\mathbf{y}) = \sum_{i=1}^{m}\{\log \alpha_i\}$. As estimativas obtidas pela maximização de $\ell(\boldsymbol{\theta}|\mathbf{y})$ são, portanto, as estimativas de Máxima Verossimilhança para o modelo TCFA definido por (3.2.1) e (3.2.2).

Por meio da equação 3.2.3, critérios de seleção de modelos podem ser desenvolvidos e aplicados para se comparar o desempenho de diferentes modelos. Os critérios mais utilizados na literatura são $AIC = -2\ell(\boldsymbol{\theta}|\mathbf{y}) + 2T$, $BIC = -2\ell(\boldsymbol{\theta}|\mathbf{y}) + Tln(n)$ e $CAIc = -2\ell(\boldsymbol{\theta}|\mathbf{y}) + T(ln(n) + 1)$, em que T denota o número de parâmetros livres do modelo.

O algoritmo EM

Como a função de log-verossimilhança envolve expressões complexas, torna-se bastante árdua a tarefa de avaliá-la analiticamente. Para modelos lineares (ou não lineares) de efeitos mistos, um algoritmo do tipo EM foi desenvolvido por Matos et al. (2013a) para se obter estimativas de máxima verossimilhança. Neste trabalho, foi proposto um algoritmo similar aplicado ao modelo TCFA, considerando $\mathbf{y}_i$ e $\mathbf{z}_i$ como dados faltantes para se obter estimativas atualizadas (passo M) para todos os parâmetros envolvidos no modelo.

Sejam $\mathbf{y} = (\mathbf{y}_1^\top, \ldots, \mathbf{y}_n^\top)^\top$, $\mathbf{z} = (\mathbf{z}_1^\top, ..., \mathbf{z}_n^\top)^\top$, $\mathbf{V} = vec(\mathbf{V}_1, \ldots, \mathbf{V}_n)$ e $\mathbf{C} = vec(\mathbf{C}_1, \ldots, \mathbf{C}_n)$ tal que se observa $(\mathbf{V}_i, \mathbf{C}_i)$ para o i-ésimo indivíduo. Para procedimento de estimação, $\mathbf{z}$, $\mathbf{Q}$ e $\mathbf{C}$ são tratados como dados faltantes e, portanto, o conjunto de dados completos é dado por $\mathbf{y}_c = (\mathbf{C}^\top, \mathbf{V}^\top, \mathbf{y}^\top, \mathbf{z}^\top)^\top$. Assim, a função de log-verossimilhança dos dados completos será $\ell_c(\boldsymbol{\theta}|\mathbf{y}_c) = \sum_{i=1}^{m} \ell_i(\boldsymbol{\theta}|\mathbf{y}_c)$, em que

$$\begin{aligned} \ell_i(\boldsymbol{\theta}|\mathbf{y}_c) &= cte - \frac{1}{2}\left[\log|\boldsymbol{\Psi}| + (\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \boldsymbol{\Lambda}\mathbf{z}_i)^\top \Psi^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \boldsymbol{\Lambda}\mathbf{z}_i) + \log|\boldsymbol{\Omega}|\right] \\ &+ \frac{1}{2}\left[\mathbf{z}_i^\top \boldsymbol{\Omega}^{-1}\mathbf{z}_i\right], \end{aligned} \tag{3.2.4}$$

e cte é uma constante que independe do vetor de parâmetros $\boldsymbol{\theta}$. Dada a atual estimativa de $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}^{(k)}$, o passo E do algoritmo resume-se ao cálculo da esperança condicional da função de log-verossimilhança dada por

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

em que

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

e

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

com $\widehat{a_i}^{(k)} = tr\left(\widehat{\mathbf{y}_i\mathbf{y}_i^\top}^{(k)}\boldsymbol{\Psi}^{-1(k)} - 2\widehat{\mathbf{y}_i\mathbf{z}_i^\top}^{(k)}\boldsymbol{\Lambda}^{\top(k)}\boldsymbol{\Psi}^{-1(k)} + \boldsymbol{\Lambda}^{(k)}\widehat{\mathbf{z}_i\mathbf{z}_i^\top}^{(k)}\boldsymbol{\Lambda}^{\top(k)}\boldsymbol{\Psi}^{-1}\right)$, $\widehat{\mathbf{z}_i\mathbf{z}_i^\top}^{(k)} = E\left[\mathbf{z}_i\mathbf{z}_i^\top|\mathbf{V}_i, \mathbf{C}_i, \boldsymbol{\theta}^{(k)}\right] = \boldsymbol{\Delta}^{(k)} + \boldsymbol{\Delta}^{(k)}\boldsymbol{\Lambda}^{\top(k)}\boldsymbol{\Psi}^{-1(k)}\widehat{b}_i^{(k)}\Psi^{-1(k)}\boldsymbol{\Lambda}^{\top(k)}\boldsymbol{\Delta}^{(k)}$,
$\widehat{b}_i^{(k)} = \left(\widehat{\mathbf{y}_i\mathbf{y}_i^\top}^{(k)} - \widehat{\mathbf{y}_i^{(k)}}\boldsymbol{\beta}^{\top(k)}\mathbf{X}_i^\top - \mathbf{X}_i\boldsymbol{\beta}\widehat{\mathbf{y}_i^{(k)}} + \mathbf{X}_i\boldsymbol{\beta}^{(k)}\boldsymbol{\beta}^{\top(k)}\mathbf{X}_i^\top\right)$ $\boldsymbol{\Delta}^{(k)} = \left(\boldsymbol{\Omega}^{-1} + \boldsymbol{\Lambda}^{\top(k)}\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}\right)$, $\widehat{\mathbf{z}}_i = E\left[\mathbf{z}_i|\mathbf{V}_i, \mathbf{C}_i, \boldsymbol{\theta}^{(k)}\right] = \boldsymbol{\Delta}^{(k)}\boldsymbol{\Lambda}^{t(k)}\boldsymbol{\Psi}^{-1(k)}\left(\widehat{\mathbf{y}_i^k} - \mathbf{X}_i\boldsymbol{\beta}^{(k)}\right)$, $\widehat{\mathbf{y}_i\mathbf{z}_i^\top} = E\left[\mathbf{y}_i\mathbf{z}_i^\top|\mathbf{V}_i, \mathbf{C}_i, \boldsymbol{\theta}^{(k)}\right] = (\widehat{\mathbf{y}_i\mathbf{y}_i^\top}^{(k)} - \widehat{\mathbf{y}_i^{(k)}}\boldsymbol{\beta}^{\top(k)}\mathbf{X}_i^\top)\boldsymbol{\Psi}^{-1(k)}\boldsymbol{\Lambda}^{(k)}\boldsymbol{\Delta}^{(k)}$.

É evidente que o passo E se simplifica por se fazer necessário somente o cálculo de $\widehat{\mathbf{y}_i\mathbf{y}_i^\top}^{(k)} = E\left[\mathbf{y}_i\mathbf{y}_i^\top|\mathbf{V}_i, \mathbf{C}_i, \boldsymbol{\theta}^{(k)}\right]$ e $\widehat{\mathbf{y}}_i^{(k)} = E\left[\mathbf{y}_i|\mathbf{V}_i, \mathbf{C}_i, \boldsymbol{\theta}^{(k)}\right]$, que representam a média e o segundo momento de uma distribuição Normal multivariada truncada. Essas expressões podem ser determinadas de forma fechada, como funções de probabilidades associadas à distribuição multivariada Normal, utilizando-se uma sequência simples de transformações. Para mais detalhes sobre o cálculo desses momentos, veja Vaida and Liu (2009).

Sendo assim, dadas as esperanças condicionais, disponíveis no passo E, o passo M pode ser obtido pela maximização condicional (CM) do seguinte sistema de equações:

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

em que $\boldsymbol{\Lambda}_j^\top$ é a j-ésima linha de $\boldsymbol{\Lambda}$, $j = 1, \ldots, p$, e $\mathbf{A}_i = \widehat{\mathbf{y}_i\mathbf{y}_i^\top} - 2\widehat{\mathbf{y}_i^{(k)}}\boldsymbol{\beta}^{\top(k)}\mathbf{X}_i^\top - 2\widehat{\mathbf{y}_i\mathbf{z}_i^\top}\boldsymbol{\Lambda}^\top + 2\boldsymbol{\Lambda}^{(k)}\mathbf{z}_i\boldsymbol{\beta}^{\top(k)}\mathbf{X}_i^\top + \mathbf{X}_i\boldsymbol{\beta}^{(k)}\boldsymbol{\beta}^{\top(k)}\mathbf{X}_i^\top + \boldsymbol{\Lambda}^{(k)}\widehat{\mathbf{z}_i\mathbf{z}_i^\top}^{(k)}\boldsymbol{\Lambda}^{\top(k)}$. Esse processo é repetido iterativamente até que algum critério de parada seja satisfeito. Usualmente, utiliza-se alguma medida de distância entre os últimos dois valores sucessivos da função log-verossimilhança $\ell(\boldsymbol{\theta}|\mathbf{y})$, tais como $|\ell(\widehat{\boldsymbol{\theta}}^{(k+1)}) - \ell(\widehat{\boldsymbol{\theta}}^{(k)})|$ ou $|\ell(\widehat{\boldsymbol{\theta}}^{(k+1)})/\ell(\widehat{\boldsymbol{\theta}}^{(k)}) - 1|$, de forma que esta seja a menor possível. Seguindo, Hughes (1999), Matos et al. (2013a) e Vaida and Liu (2009), a variância dos efeitos fixos do modelo TCFA pode ser obtida por

$$
Var(\widehat{\boldsymbol{\beta}}) = \left(\sum_{i=1}^{m} \mathbf{X}_i^\top \boldsymbol{\Sigma}_i^{-1}\mathbf{X}_i - \mathbf{X}_i^\top \boldsymbol{\Sigma}_i^{-1} Var(\mathbf{y}_i|\mathbf{V}_i, \mathbf{C}_i)\boldsymbol{\Sigma}_i^{-1}\mathbf{X}_i\right)^{-1}. \tag{3.2.5}
$$

Na próxima seção, será proposta a modelagem via inferência Bayesiana para o modelo TCFA.

3.2.2 Estimação Bayesiana

Nesta seção, propomos uma modelagem hierárquica Bayesiana para o modelo TCFA, métodos de seleção de modelos via inferência Bayesiana, bem como métodos de diagnóstico de dados influentes baseados na medida de divergência q (Peng and Dey, 1995). Primeiro, serão definidas as distribuições para os parâmetros e, então, serão apresentadas as distribuições condicionais completas, que possibilitam o uso de técnicas MCMC, tal como o amostrador de Gibbs.

Distribuições *a priori* e *a posteriori*

Considerando as especificações do modelo dadas em (3.2.1) e (3.2.2), o modelo de análise confirmatória Tobit pode ser facilmente representado como:

$$
\begin{aligned}
\mathbf{y}_i|\mathbf{z}_i, \mathbf{C}_i, \mathbf{V}_i, \boldsymbol{\theta} &\sim TN(\mathbf{x}_{ij}^\top\boldsymbol{\beta} + \boldsymbol{\Lambda}_r^\top\mathbf{z}_i, \Psi jj; (-\infty, \mathbf{q}_i)) \qquad (3.2.6)\\
\mathbf{z}_i|\mathbf{C}_i, \mathbf{V}_i &\sim N_q(\mathbf{0}, \boldsymbol{\Omega}),
\end{aligned}
$$

em que $\boldsymbol{\Lambda} = (\boldsymbol{\Lambda}_1, \ldots, \boldsymbol{\Lambda}_q)_{p\times q}$ e $\boldsymbol{\Lambda}_r$ representa a r-ésima coluna de $\boldsymbol{\Lambda}$, $r = 1, \ldots, q$, $i = 1, \ldots, n$, $j = 1, \ldots, p$ e $TN(\mu, \sigma^2; a)$ denota a distribuição normal truncada no ponto a e $q_{ij} = 0$ se $C_{ij} = 1$ (caso censurado) ou $q_{ij} = \infty$ se $C_{ij} = \mathbf{0}$ (caso não censurado).

Para especificação completa sob ponto de vista Bayesiano, faz-se necessário obter as distribuições *a priori* dos parâmetros $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \boldsymbol{\Lambda}, \boldsymbol{\Psi})$. Vale ressaltar que, tal como na análise frequentista, será assumindo aqui também que $\boldsymbol{\Omega}$ é uma matriz conhecida. Em modelos lineares mistos, é comum a escolha de distribuições *a priori* condicionalmente conjugadas para se assegurar que a distribuição *a posteriori* seja própria, veja Hobert and Casella (1996). Para o modelo TCFA, foram utilizadas também distribuições *a priori* próprias condicionalmente conjugadas tal que

$$\boldsymbol{\beta} \sim N_p(\boldsymbol{\beta}_{\mathbf{0}}, \boldsymbol{S}_{\boldsymbol{\beta}}), \tag{3.2.7}$$

$$\boldsymbol{\Lambda}_r \sim N_p(\boldsymbol{\Lambda}_{\mathbf{0}}, \boldsymbol{S}_{\boldsymbol{\Lambda}}), r = 1, \ldots, q, \tag{3.2.8}$$

$$\Psi_{jj} \sim IGamma(k_0/2, v_0/2), \tag{3.2.9}$$

em que $IGamma(a, b)$ é a distribuição Gama inversa com média $b/(a-1)$, $a > 1$, e Ψ_{jj} é o j-ésimo elemento da matriz diagonal $\boldsymbol{\Psi}$, $j = 1, \ldots, p$. Assim, o núcleo da densidade a posteriori é dado por:

$$\begin{aligned}\pi(\boldsymbol{\theta}|\mathbf{y}, \mathbf{z}, \mathbf{C}, \mathbf{V}) &\propto \textstyle\prod_{i=1}^n TN(\mathbf{y}_i|\mathbf{z}_i, \mathbf{C}_i, \mathbf{V}_i, \boldsymbol{\theta}) \prod_{r=1}^q N_p(\boldsymbol{\Lambda}_r|\boldsymbol{\Lambda}_{\mathbf{0}}, \boldsymbol{S}_{\boldsymbol{\Lambda}}) \times \\ &\textstyle\prod_{j=1}^p [N_p(\beta_j|\boldsymbol{\beta}_{\mathbf{0}}, \boldsymbol{S}_{\boldsymbol{\beta}}) IGamma(\Psi_{jj}|k_0, v_0)].\end{aligned} \tag{3.2.10}$$

De (3.2.6), tem-se que $\mathbf{z}_i|\mathbf{y}_i, \mathbf{C}_i, \mathbf{V}_i, \boldsymbol{\theta} \sim N_q(\widehat{\mathbf{z}}_i, \boldsymbol{\Delta})$, em que $\widehat{\mathbf{z}}_i = \boldsymbol{\Delta}\boldsymbol{\Lambda}^\top\boldsymbol{\Psi}^{(-1)}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})$ e $\boldsymbol{\Delta} = (\boldsymbol{\Omega}^{-1} + \boldsymbol{\Lambda}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda})^{-1}$. Como a distribuição *a posteriori* apresentada na (3.2.10), não é tratável analiticamente, algoritmos do tipo MCMC podem ser empregados para se obter aproximações numéricas para as distribuições marginais *a posteriori*. Sejam $\boldsymbol{\theta}_1|\mathbf{y}, \mathbf{C}, \mathbf{V}, \boldsymbol{\theta}_{(-\boldsymbol{\theta}_1)}$ a densidade condicional completa de $\boldsymbol{\theta}_1$ e $\boldsymbol{\theta}_1|\mathbf{y}, \mathbf{C}, \mathbf{V}, \boldsymbol{\theta}_{(-\boldsymbol{\theta}_1)} = \boldsymbol{\theta}_1|\mathbf{y}, \boldsymbol{\theta}_{(-\boldsymbol{\theta}_1)}$. Assim, tem-se:

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

em que $\mathbf{A}_\beta = (\mathbf{S}_\beta^{-1} + \sum_{i=1}^m \mathbf{X}_i^\top\boldsymbol{\Psi}^{-1}\mathbf{X}_i)^{-1}$, $\boldsymbol{u}_{\boldsymbol{\beta}} = \boldsymbol{(\beta_0^\top S_\beta^{-1} + \sum_{i=1}^m (y_i - X_i^\top \beta_0)^\top \Psi^{-1} X_i^\top)}$, $s = \sum_{i=1}^m(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \Lambda_r^\top\mathbf{z}_i)$, $\mathbf{A}_\Lambda = (\sum_{i=1}^m \mathbf{z}_i\mathbf{z}_i^\top\boldsymbol{\Psi}^{-1} + \boldsymbol{\Lambda}_0^{-1})^{-1}$, $\boldsymbol{u}_{\boldsymbol{\Lambda}} = \boldsymbol{\sum_{i=1}^m A_{ir}^\top \Psi^{-1} z_{ir} + \Lambda_0^\top S_\Lambda^{-1}}$, $\mathbf{A}_{ir} = \mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \sum_{l\neq r}^{p_{l=1}} \Lambda_l\mathbf{z}_{il}$, $r = 1, ..., q$.

Na subseção subsequente, serão discutidos os critérios de comparação de modelos. Para maiores detalhes, veja Lachos et al. (2013b).

Critérios Bayesianos de seleção de modelos

Neste trabalho será utilizada a medida de predição condicional ordinal (em inglês, *conditional predictive ordinate - CPO*), um dos critérios de comparação de modelos mais difundidas no contexto Bayesiano. Essa medida é derivada da distribuição preditiva a posteriori (veja Carlin and Louis, 2008). Seja $\mathcal{D}$ o conjunto completo de dados e $\mathcal{D}^{(-i)}$ o conjunto de dados sem a i-ésima observação. Denote a densidade *a posteriori* de θ dado $\mathcal{D}^{(-i)}$ por $\pi(\theta|\mathcal{D}^{(-i)})$. Para a i-ésima observação, a medida CPO_i pode ser escrita por $CPO_i = \int_\Theta f(\mathbf{y}_i|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{\mathcal{D}}^{(-i)})d\boldsymbol{\theta} = \left\{\int_\Theta \frac{\pi(\boldsymbol{\theta}|\boldsymbol{\mathcal{D}})}{f(\mathbf{y}_i|\boldsymbol{\theta})}d\boldsymbol{\theta}\right\}^{-1}$. Para o modelo TCFA, uma forma fechada de CPO_i não é possível analiticamente. Contudo, uma estimativa via Monte Carlo de CPO_i pode ser obtida ao se utilizar uma única amostra MCMC da distribuição *a posteriori* $\pi(\boldsymbol{\theta}|\mathcal{D})$ sob a aproximação utilizando uma média harmônica, veja, por exemplo, Dey et al. (1997), tal que $\widehat{CPO}_i = \left\{\frac{1}{Q}\sum_{q=1}^{Q}\frac{1}{f(\mathbf{y}_i|\boldsymbol{\theta}_q)}\right\}^{-1}$, em que $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_Q$ é uma amostra MCMC válida de tamanho Q (ou seja, após o descarte das primeiras estimativas - *burn-in* - e selecionando sistematicamente observações de forma espaçadas) de $\pi(\theta|\mathcal{D})$. Um resumo da medida CPO_i é a função de log-verossimilhança pseudo-marginal (LPML), definida por $LPML = \sum_{i=1}^{n}\log(\widehat{CPO}_i)$. Quanto maior é o valor de $LPML$, melhor a qualidade do ajuste do modelo.

Outros critérios de comparação de modelos que também podem ser utilizados são: o de informação dos desvios (DIC) proposto por Spiegelhalter et al. (2002), o critério de Akaike esperado (EAIC) e o da informação Bayesiana esperada (EBIC) tal como apresentados em Carlin and Louis (2001). Essas medidas se baseiam na média *a posteriori* da função desvio, que pode ser aproximada por $\overline{D} = \sum_{q=1}^{Q} D(\boldsymbol{\theta}_q)/Q$, em que $D(\boldsymbol{\theta}) = -2\sum_{i=1}^{n}\log\left[f(\mathbf{y}_i|\theta)\right]$. O critério DIC pode ser estimado utilizando os resultados via MCMC como $\widehat{DIC} = \overline{D} + \widehat{\rho_D}$, em que $\rho_D = E\{D(\boldsymbol{\theta})\} - D\{E(\boldsymbol{\theta})\}$ é o número efetivo de parâmetros e $D\{E(\boldsymbol{\theta})\}$ é a função desvio avaliada na média *a posteriori*. De forma parecida, os critérios EAIC e EBIC podem ser estimados como: $\widehat{EAIC} = \overline{D} + 2\#(\vartheta)$ e $\widehat{EBIC} = \overline{D} + \#(\vartheta)\log(n)$, em que $\#(\vartheta)$ é o número de parâmetros do modelo. Observe que para todos esses critérios, a estimação da função de verossimilhança dada em (3.2.3) é imprescindível. Contudo, ela pode ser facilmente obtida segundo a abordagem proposta, tratando separadamente as componentes associadas aos dados observados e censurados de $\mathbf{y}_i$ como apresentado na Subseção 3.2.1.

Medidas de diagnóstico Bayesiano para dados influentes

Inferências em modelos de análise fatorial com covariáveis, assim como em modelos de regressão, são fortemente afetadas pela inclusão ou exclusão de um conjunto de observações. Para se estudar os efeitos de observações influentes na análise, esquemas de perturbação do modelo têm sido desenvolvidos na literatura estatística (veja Cook (1986)). Os esquemas mais comuns se baseiam no método de exclusão de casos (Cook and Weisberg, 1982), em que o efeito de se retirar todos os casos influentes da análise é avaliado. Para o modelo TCFA, utilizou-se também um esquema de exclusão de casos sob perspectiva Bayesiana tendo por base o uso de funções de perturbação.

Funções de perturbação foram introduzidas por Kass et al. (1989) e Weiss (1996). Por meio dessas funções, faz-se possível a avaliação da sensibilidade dos pressupostos do modelo M na distribuição *a posteriori* $\pi(\boldsymbol{\theta}|\mathbf{y}, M)$. Suponha que $\pi(\boldsymbol{\theta}|\mathbf{y}, M_1)$ seja a distribuição *a posteriori* de $\boldsymbol{\theta}$ sob o modelo M_1 e $\pi(\boldsymbol{\theta}|\mathbf{y}, M_2)$ é sua distribuição *a posteriori* sob o modelo M_2. Assim, a função de perturbação por exclusão de casos é definida por $p(\boldsymbol{\theta}) = \frac{\pi(\boldsymbol{\theta}|\mathbf{y},M_2)}{\pi(\boldsymbol{\theta}|\mathbf{y},M_1)}$.

Considerando um subconjunto I com k elementos do conjunto $\{1, \ldots, n\}$. Quando o subconjunto I é deletado dos dados $\mathbf{y}$, denote esses dados eliminados por $\mathbf{y}_I$ e $\mathbf{y}_{(-I)}$ os dados restantes. Logo, a funcão de perturbação por exclusão de casos é $p(\boldsymbol{\theta}) = \pi\left(\boldsymbol{\theta}|\mathbf{y}_{(-I)}\right) / \pi\left(\boldsymbol{\theta}|\mathbf{y}\right)$. Após algumas manipulações algébricas, a função de perturbação pode ser definida por

$$p(\boldsymbol{\theta}) = \frac{\left[\prod_{i \in I} f(\mathbf{y}_i|\boldsymbol{\theta})\right]^{-1}}{E_{\boldsymbol{\theta}|\mathbf{y}}\left\{\left[\prod_{i \in I} f(\mathbf{y}_i|\boldsymbol{\theta})\right]^{-1}\right\}}, \tag{3.2.11}$$

em que $f(\mathbf{y}_i|\boldsymbol{\theta})$ representa a função de verossimilhança dada na Equação (3.2.3).

A função de perturbação dos parâmetros do modelo TCFA por exclusão de casos pode ser aproximada pelo uso da distribuição marginal de $\mathbf{y}$ e técnicas de MCMC pela amostragem da distribuição *a posteriori*. De fato, quando o subconjunto $I = \{i\}$ é considerado na análise e $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_Q$ é uma amostra MCMC válida de tamanho Q de $\pi(\boldsymbol{\theta} \mid \mathbf{y})$, a aproximação via Monte Carlo da funcão de perturbação $p(\boldsymbol{\theta})$ é dada por

$$\widehat{p(\boldsymbol{\theta})} = \widehat{\mathrm{CPO}_i}\left[\phi_{p^o}(\mathbf{y}_i^o; \mathbf{X}_i^o\boldsymbol{\beta}, \boldsymbol{\Sigma}_i^{oo})\Phi_{p^c}(\mathbf{V}_i^c; u_i.\mathbf{S}_i)\right]^{-1}, \tag{3.2.12}$$

Outra abordagem usual para se quantificar as observações influentes é utilizar medidas de divergência entre distribuições *a posteriori* com e sem um determinado subconjunto de dados. A medida de divergência q entre duas densidades π_1 e π_2 para $\boldsymbol{\theta}$ foi definida por Csiszár (1967) como

$$d_q(\pi_1, \pi_2) = \int q\left(\frac{\pi_1(\boldsymbol{\theta})}{\pi_2(\boldsymbol{\theta})}\right)\pi_2(\boldsymbol{\theta})d\boldsymbol{\theta}, \tag{3.2.13}$$

em que $q(.)$ é uma função convexa tal que $q(1) = 0$. Medidas específicas de divergência são obtidas ao se considerar certas funções $q(\cdot)$. Por exemplo, a medida de divergência Kullback-Leibler é obtida quando $q(z) = -\log(z)$, a medida de distância J (uma versão simétria da divergência de Kullback-Leibler) é obtida quando $q(z) = (z-1)\log(z)$ e a distância L_1 se define por $q(z) = |z-1|$ (Lachos et al., 2013b).

A medida de influência q dos dados $\mathbf{y}_I$ na distribuição *a posteriori* de $\boldsymbol{\theta}$, $d_q(I) = d_q(\pi_1, \pi_2)$, é obtida ao considerar $\pi_1(\boldsymbol{\theta}) = \pi_1(\boldsymbol{\theta}|\mathbf{y}_{(-I)})$ e $\pi_2(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta}|\mathbf{y})$ na Equação (3.2.13), e pode ser escrita como

$$d_q(I) = E_{\boldsymbol{\theta}|\mathbf{y}}\left[q(p(\boldsymbol{\theta}))\right], \tag{3.2.14}$$

em que os valores esperados são obtidos com relação à distribuição *a posteriori* dos dados não perturbados. Essas medidas de influência têm sido utilizadas por Peng and Dey (1995), Weiss (1996) e mais recentemente por Vidal and Castro (2010).

Cabe ressaltar que a medida de influência $d_q(I)$ por si mesma não determina se uma observação é influente ou não. Para tanto, faz-se necessária a definição de um ponto de corte que possibilita determinar se um pequeno subconjunto de observações é influente. Nesse contexto, a proposta dada por Peng and Dey (1995) foi utilizada. Considere a função de probabilidade de uma moeda viesada, que é dada por $\pi_1(x \mid p) = p^x(1-p)^{1-x}$, com $x = 0, 1$, enquanto a função de probabilidade de uma moeda não viciada é dada por $\pi_2(x \mid p) = 0.5$. De (3.2.13), a medida de divergência q entre uma moeda viciada e outra não viciada é dada por

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

em que $d_q(p)$ cresce quando p se afasta de 0.5, é simétrica em torno de $p = 0.5$ e atinge seu valor mínimo em $p = 0.5$. Adicionalmente, se $d_q(0.5) = 0$ então $\pi_1 = \pi_2$. Consequentemente, se for considerado que $p \geq 0.85$ (ou $p \leq 0.15$) representa um forte viés de uma moeda, então, $d_{L_1}(0.85) = 0.70$ e, portanto, pode-se dizer que uma observação é influente se $d_{L_1}(i) \geq 0.70$, $i = 1, \ldots, m$. De forma similar, para a medida de divergência de Kullback-Leibler, tem-se $d_{KL}(0.85) = 0.33$ e para a medida de distância J, $d_J(0.85) = 0.61$. Esses pontos de corte foram utilizados nos estudos empíricos.

3.2.3 Aplicações

Estudos de simulação e uma aplicação a dados reais foram utilizados para avaliar os métodos propostos. A seguir, será apresentado um resumo dos resultados encontrados.

Estudos de simulação

Na parte da aplicação, dois estudos de simulação foram realizados. O primeiro teve por objetivo comparar as estimativas obtidas de um modelo de análise fatorial confirmatória clássica (CFA) e de um modelo de análise fatorial confirmatória Tobit (TCFA) para diferentes percentuais de censura na amostra e diferentes abordagens (clássica ou Bayesiana). Já o segundo estudo, a finalidade principal era avaliar se as estimativas obtidas via o algoritmo EM proposto apresentam boas propriedades assintóticas.

Para tanto, os dados foram simulados do modelo TCFA definido em (3.2.1) e (3.2.2), com $p = 5$, $k = 3$ e $q = 2$. Os valores verdadeiros dos parâmetros $\boldsymbol{\beta}$, $\boldsymbol{\Lambda}$ e $\boldsymbol{\Psi}$ foram:

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

Como uma análise confirmatória, a matriz $\boldsymbol{\Omega}$ foi fixada previamente em $0.6\mathbf{I} + 0.4\mathbf{J}$, em que $\mathbf{I}$ é uma matriz identidade (2×2) e $\mathbf{J}$ é uma matriz (2×2) de uns. Os valores iniciais fixados para ambas abordagens, clássica e Bayesiana, foram:

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

A matriz de covariáveis $\mathbf{X}_i$ foi gerada pelo produto *Kronecker* de duas matrizes A e B, em que A é uma matriz $(1 \times p)$ de uns e B é um vetor transposto associado ao $\mathbf{x}_{ij}$, $j = 1, 2, 3$. A covariável x_{i1} foi igual a um para todo $i = 1, \ldots, n$ (intercepto); x_{i2} foi gerada de uma distribuição Normal com média 6 e variância 1 e x_{i3} foi gerada independentemente de uma distribuição Bernoulli(0.5).

Todos os programas foram implementados no *software R*. Em todas as situações, foram comparadas as consequências na inferência sobre os parâmetros quando o mecanismo de censura é levado em consideração (modelo TCFA) ou ignorado (modelo CFA).

Estudo 1: Este estudo teve por objetivo a investigação do impacto nas estimativas dos modelos TCFA e CFA para diferentes percentuais de censura na amostra. Para tanto, o algoritmo do tipo EM proposto foi utilizado para se obter estimativas por máxima verossimilhança e duas cadeias MCMC independentes de tamanho 20.000 foram geradas na abordagem Bayesiana e as seguintes distribuições *a priori* foram designadas para os parâmetros do modelo: $\beta_l \sim N_1(1, 10^2)$, $\Lambda_{1k} \sim N_1(-0.8, 1), k = 1, 2, 3$, $\Lambda_{1k} \sim N_1(0, 10^{-6}), k = 3, 4$, $\Lambda_{2k} \sim N_1(0, 10^{-6})$, $k = 1, 2, 3$, $\Lambda_{2k} \sim N_1(0.8, 1)$ $k = 3, 4$ e $\Psi_l \sim IGamma(2, 1)$. Por meio da Tabela 3.1, os resultados para ambas abordagens (clássica e Bayesiana) indicaram que o modelo TCFA foi o que melhor se ajustou aos dados independente da quantidade de casos censurados na amostra. Dado que as observações foram geradas deste modelo, este estudo evidenciou ainda mais a necessidade de se considerar a informação sobre o procedimento de censura na análise, uma vez que as estimativas obtidas pelo modelo CFA foram as que mais se distanciaram dos seus valores verdadeiros. Ao compararmos as abordagens frequentistas e Bayesianas, verificou-se que ambas obtiveram resultados similares. A convergência das cadeias MCMC foi alcançada segundo as análises dos gráficos das trajetórias das amostras da cadeia (*trace plots*), gráficos de auto correlação e das medidas de diagnóstico Gelman-Rubin $\hat{R}$.

Estudo 2: Para se estudar algumas propriedades assintóticas associadas às estimativas por máxima verossimilhança, cinco tamanhos de amostras foram fixados ($n = 30, 50, 150, 350$ e 600) e medidas de vício e erro quadrático médio (EQM) foram avaliadas. Para cada tamanho de amostra, 100 réplicas de dados simulados do modelo TCFA foram geradas e, utilizando o algoritmo do tipo EM proposto, estimativas sob a modelagem via CFA e TCFA foram obtidas. Para este estudo, o percentual de censura da amostra foi de 20%. Os resultados indicaram que para ambos modelos as estimativas obtidas pelo algoritmo proposto apresentaram boas propriedades assintóticas. Como esperado, as medidas de vício e EQM se aproximaram de zero à medida que o tamanho da amostra crescia. Observou-se também que o modelo TCFA foi mais estável e apresentou melhor desempenho que o modelo que ignora censura.

Aplicação a dados reais

O banco de dados utilizado nesta aplicação refere-se ao desempenho de 502 alunos no teste de leitura EGRA *(Early Grade Reading Assessment)* desenvolvido pela agência norte-americana USAID (*U.S. Agency for International Development*). Este exame é um instrumento para medir o progresso de aprendizado em leitura de alunos em séries iniciais (RTI International, 2009). Mais especificamente, o teste EGRA avalia quão bem as crianças nas séries iniciais estão evoluindo no aprendizado de leitura e também determina em quais áreas desse processo de desenvolvimento precisa ser melhorado. Este teste é aplicado a alunos da América Latina e no Caribe e é administrado oralmente, um estudante por vez. Para este estudo, foram utilizados os resultados de quatro tarefas/subtestes do teste EGRA de língua Espanhola aplicados em 2007 para estudantes do Peru. Esses

Tabela 3.1: **Estudo 1.** MC mean e MC Sd são, respectivamente, as estimativas médias e os desvios-padrão via Monte Carlo de 100 réplicas. Os valores em parênteses são as medidas de erro-padrão segundo a Equação 3.2.5.

		Estimativas EM				Estimativas bayesianas			
Percentual de censura	Parâmetros	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

subtestes se referem a: (1) habilidade de reconhecer letras do alfabeto; (2) habilidade de reconhecer palavras simples; (3) simples decodificação de palavras sem sentido; e (4) leitura de uma passagem.

O desempenho dos estudantes em cada subteste foi calculado por uma medida de velocidade dada por: $\text{Velocidade}_{ij} = \frac{Y_{ij}}{Tempo_{ij}}$, em que: Y_{ij}=número de letras/palavras lidas pelo estudante i na tarefa j em até 60 segundos e $Tempo_{ij}$= tempo (em segundos) gasto pelo estudante i na tarefa j (menor ou igual a 60). Essa medida indica que, quanto maior é a fluência do estudante em Língua Espanhola, maior será sua medida de velocidade com relação às tarefas avaliadas. A aplicação de técnicas de análise fatorial para este estudo tem por objetivo verificar que, dada a alta associação entre essas medidas de velocidade, a redução da dimensionalidade pode ser justificada e uma medida geral de habilidade (interpretada como Fluência em Língua Espanhola) pode melhor representar os dados. Contudo, devido à natureza do teste que apresenta restrições quanto ao tempo de resposta em cada tarefa, baixos escores nos subtestes não representam necessariamente baixas medidas de fluência na língua avaliada. Sob a suposição de que o tempo associado às tarefas avaliadas não foi suficiente para estimar bem a fluência para indíviduos com baixos escores, as 10% menores medidas de velocidade foram consideradas como casos censurados. Utilizando esta definição de observações censuradas, o modelo TCFA foi definido por:

$$\mathbf{Velocidade}_{y_i} = \mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\Lambda}\mathbf{Z}_i + \boldsymbol{\epsilon}_i, \tag{3.2.15}$$

em que $\mathbf{Velocidade}_{y_i} = (Velocidad_{y_i1}, \ldots, Velocidad_{y_i4})^\top$ é um vetor de dimensão (4×1) das medidas de velocidade do estudante i nas quatros tarefas, $i = 1, \ldots, 502$; $\mathbf{X}_i$ é a matriz de delineamento de dimensão (4×5) correspondendo aos seguintes efeitos fixos: $\beta_1 =$sexo (0=Feminino, 1=Masculino); $\beta_2 =$série escolar (0=Segundo ano; 1=Terceiro ano); $\beta_3 =$zona de residência (0=Rural, 1=Urbana);$\beta_4 =$idade; $\boldsymbol{\Lambda}$ é um vetor (4×1) das cargas fatoriais; $\mathbf{Z}_i$ é o fator latente associado à habilidade geral (fluência em Espanhol); $\boldsymbol{\epsilon}_i$ é um vetor (4×1) dos erros aleatórios e Ω é um escalar, fixado em 1.

Na base de dados, observa-se que 157 estudantes são do sexo feminino e 345 do sexo masculino; 354 são da segunda série e 148, da terceira; 250 são provenientes da zona urbana e 252 da zona rural; 51% dos estudantes tinha até sete anos de idade. Como nos estudos simulados, foram obtidas estimativas via máxima verossimilhança e análise Bayesiana incluindo ou não a informação sobre censura. As variáveis respostas foram padronizadas para se evitar o impacto nas estimativas devido à diferença nas escalas dessas variáveis.

Para análise Bayesiana, duas cadeias MCMC paralelas e independentes de tamanho 50 000 foram geradas, as primeiras 5 000 iterações de cada foram descartadas (procedimento conhecido como *burn-in*) e para eliminar problemas associados às auto correlações das amostras, as estimativas consideradas foram obtidas considerando um sistema de amostragem sistemática (*lag* igual a 20). Seguindo Lopes and West (2004), foram consideradas as seguintes distribuições *a priori* independentes: $\beta \sim N(0, 10), \Lambda \sim N(0, 1)\mathbf{1}(\lambda_i > 0)$ e $\Psi \sim IGamma(1.1, 0.05)$. A convergência das cadeias MCMC foi monitorada utilizando gráficos das trajetórias (*trace plots*), das auto correlações (ACF) e as medidas de diagnóstico de Gelman-Rubin $\hat{R}$.

Por meio da Tabela 3.2, verificou-se que ambas abordagens (frequentista e Bayesiana) apresentaram estimativas próximas. Apesar da padronização das variáveis, observou-se que garotos mais jovens da terceira série provenientes da zona urbana são os estudantes com medidas de velocidade mais altas. Todas as variáveis explicativas incluídas nos modelos foram significativas. Com relação ao desempenho dos modelos CFA e TCFA, verificou-se que as estimativas do vetor de cargas fatoriais, bem como as variâncias específicas (matriz $\boldsymbol{\Psi}$), foram levemente maiores para o modelo em que a informação sobre os dados censurados é levada em consideração. Altos valores para as cargas fatoriais confirmam nossa hipótese inicial de que um fator latente geral associado à fluência em Língua Espanhola é adequada. A variável resposta de menor impacto na estimação desse fator latente e alta variabilidade ($\Psi_{44} = 0.788$) foi o subteste 1 (reconhecer letras do alfabeto). Segundo os critérios

Tabela 3.2: Dados EGRA. Estimativas dos parâmetros via algoritmo EM e pela análise Bayesiana para os modelos TCFA e CFA.

Parâmetros	Algoritmo EM		Análise Bayesiana						
	TCFA	CFA		TCFA			CFA		
	Estimativas	Estimativas		Média	Sd	IC(95%)	Média	Sd	IC(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)
Critérios de comparação de modelos									
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	

de comparação de modelos, a abordagem via TCFA é a preferida. Este fato corrobora os estudos simulados sobre a importância de se incluir a informação sobre as observações censuradas na análise dos dados.

Para esse banco de dados também foi feita uma análise de diagnóstico via medidas de divergência d_q para a detecção de observações influentes. Três observações foram apontadas como atípicas: #244, #316 e #289. Essas observações se referem a estudantes do sexo feminino, zona urbana que apresentaram desempenho incoerente com o esperado. A estudante codificada como #244, por exemplo, teve bom desempenho nas tarefas mais complicadas para a grande maioria dos estudantes (Subtestes 1 e 3, que correspondem, respectivamente, a habilidade de reconhecer letras do alfabeto e a habilidade simples decodificação de palavras sem sentido) e baixo desempenho nas tarefas consideradas mais fáceis para os estudantes. Retirando as observações mais atípicas (#244 e #316), realizaram-se novas análises sob o modelo TCFA para se avaliar o impacto da retirada desses dados na inferência sobre os parâmetros do modelo, contudo, observou-se que para ambas perspectivas (clássica e Bayesiana) a significância dos parâmetros, bem como o sinal dos coeficientes estimados, permaneceram os mesmos após a eliminição desses dados.

Motivados pela preocupação da sensibilidade das inferências sob presença de potenciais pontos discrepantes ou com dados provenientes de distribuições de caudas mais pesadas que a Normal, a modelagem da análise fatorial confirmatória com covariáveis utilizando a distribuição t de Student foi proposta e será apresentada a seguir.

3.3 Modelo t-TCFA

Zhou and Tan (2010) desenvolveram uma análise hierárquica robusta para modelos de análise fatorial na qual os efeitos aleatórios e os erros associados aos indivíduos tem distribuição multivariada t de Student (t-CFA). Na formulação deles, é assumido que a variável aleatória de mistura U compartilha os mesmos parâmetros para as duas fontes de variabilidade do modelo, ou seja, eles possuem os mesmos graus de liberdade, ν. Para se obter estimativas robustas para os parâmetros, neste trabalho procedemos como Zhou and Tan (2010) (veja também, Matos et al., 2013b) por considerar uma generalização do modelo Gaussiano TCFA definido em (3.2.1) e (3.2.2), com

$$(\mathbf{z}_i, \boldsymbol{\epsilon}_i)^\top \sim t_{p+q}\{\mathbf{0}, \text{Diag}(\boldsymbol{\Omega}, \boldsymbol{\Psi}), \nu\},\ i = 1, \ldots, m, \tag{3.3.1}$$

ou equivalente a: $\mathbf{z}_i \overset{iid}{\sim} t_q(\mathbf{0}, \boldsymbol{\Omega}, \nu)$ e $\boldsymbol{\epsilon}_i \overset{iid}{\sim} t_p(\mathbf{0}, \Psi, \nu)$, em que $\text{Diag}(\mathbf{A}, \mathbf{B})$ é uma matriz bloco diagonal cujos elementos pertencem às matrizes quadradas $\mathbf{A}$ e $\mathbf{B}$. Note que $\mathbf{z}_i$ e $\boldsymbol{\epsilon}_i$ são não correlacionados dado que $\text{Cov}(\mathbf{z}_i, \boldsymbol{\epsilon}_i) = E[\mathbf{z}_i \boldsymbol{\epsilon}_i^\top] = E[E(\mathbf{z}_i \boldsymbol{\epsilon}_i^\top | U_i)] = \mathbf{0}$, em que U_i é a variável de mistura gerada de Gamma$(\nu/2, \nu/2)$.

Como no caso Gaussiano, inferências clássicas para o vetor de parâmetros $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \boldsymbol{\Lambda}, \Psi)^\top$ se baseiam na distribuição marginal de $\mathbf{y}_i$ que, para os dados completos é $\mathbf{y}_i \overset{\text{ind.}}{\sim} t_p(\mathbf{X}_i \boldsymbol{\beta}, \boldsymbol{\Sigma}, \nu)$, $i = 1, \ldots, m$, com $\boldsymbol{\Sigma} = \boldsymbol{\Lambda} \boldsymbol{\Omega} \boldsymbol{\Lambda}^\top + \boldsymbol{\Psi}$. É bem conhecido na literatura estatística que estimativas obtidas de modelos utilizando a distribuição t de Student são mais robustas a observações discrepantes que aquelas de modelos que utilizam a distribução Gaussiana. Em um estudo simulado, Zhang et al. (2013) mostraram que um modelo t-CFA sem covariáveis teve desempenho substancialmente superior à análise fatorial clássica quando *outliers* estão presentes nos dados. Essa problemática também foi discutida por Wu (2010) e Matos et al. (2013b) no contexto de dados censurados em modelos lineares de efeito misto.

Como no caso normal (Seção 3.2), a função de verossimilhança pode ser avaliada tratando separamente uma componente associada aos dados observados e outras, aos dados censurados do vetor $\mathbf{y}_i$. Por meio da Proposição 1 apresentada no Apêndice C, tem-se que $\boldsymbol{\mu}_i^{co} = \mathbf{X}_i^c \boldsymbol{\beta} + \boldsymbol{\Sigma}_i^{co} (\boldsymbol{\Sigma}_i^{oo})^{-1} (\mathbf{y}_i^o - \mathbf{X}_i^o \boldsymbol{\beta})$ e $\mathbf{S}_i = \left(\frac{\nu + \delta_i^o}{\nu + p_i^o}\right) \boldsymbol{\Sigma}_i^{cc.o}$, com $\boldsymbol{\Sigma}_i^{cc.o} = \boldsymbol{\Sigma}_i^{cc} - \boldsymbol{\Sigma}_i^{co} (\boldsymbol{\Sigma}_i^{oo})^{-1} \boldsymbol{\Sigma}_i^{oc}$ e $\delta_i^o = (\mathbf{y}_i^o - \mathbf{X}_i^o \boldsymbol{\beta})^\top (\boldsymbol{\Sigma}_i^{oo})^{-1} (\mathbf{y}_i^o - \mathbf{X}_i^o \boldsymbol{\beta})$. Assim, a função de verossimilhança associada ao indivíduo i é dada por

$$L_i(\boldsymbol{\theta}) = t_{p_i^o}(\mathbf{y}_i^o; \mathbf{X}_i^o \boldsymbol{\beta}, \boldsymbol{\Sigma}_i^{oo}, \nu) T_{p_i^c}(\mathbf{V}_i^c; \boldsymbol{\mu}_i^{co}, \mathbf{S}_i^{co}, \nu + p_i^o). \tag{3.3.2}$$

A função log-verossimilhança para os dados observados ($\ell(\boldsymbol{\theta}|\mathbf{y}) = \sum_{i=1}^m \{\log L_i(\boldsymbol{\theta})\}$) pode ser obtida para cada passo do algoritmo ECM sem nenhum custo computacional, já que a cada passo E estimativas de L_i são calculadas. Nesse contexto pode-se também utilizar a função log-verossimilhança para se avaliar a convergência do algoritmo, bem como para se calcular medidas de comparação de modelos tais como o AIC, BIC e CAIc.

Com relação à estimação dos graus de liberdade, Lucas (1997) realizou um estudo sobre aspectos robustos associados ao estimador M com distribuição t de Student no caso univariado fazendo uso de funções de influência e apontou que a robustez do modelo contra observações discrepantes só é preservada se os graus de liberdade são fixados previamente. Desta forma, neste trabalho assumimos o parâmetro ν como conhecido e, para o estudo empírico, obtemo-no avaliando o valor que atinge o máximo da função log-verossimilhança para diferentes valores de graus de liberdade e, consequentemente, o valor de ν em que o modelo apresenta melhor ajuste segundo critérios AIC ou BIC (veja Lange et al., 1989; Meza et al., 2012).

3.3.1 O algoritmo EM

Sejam $\mathbf{y} = (\mathbf{y}_1^\top, \ldots, \mathbf{y}_n^\top)^\top$, $\mathbf{z} = (\mathbf{z}_1^\top, \ldots, \mathbf{z}_n^\top)^\top$, $\mathbf{V} = vec(\mathbf{V}_1, \ldots, \mathbf{V}_n)$ e $\mathbf{C} = vec(\mathbf{C}_1, \ldots, \mathbf{C}_n)$, tal que se observe $(\mathbf{V}_i, \mathbf{C}_i)$ para o i-ésimo indivíduo. Tratando $\mathbf{z}$, $\mathbf{Q}$ e $\mathbf{C}$ como dados faltantes e o conjunto de dados completos como $\mathbf{y}_c = (\mathbf{C}^\top, \mathbf{V}^\top, \mathbf{y}^\top, \mathbf{z}^\top)^\top$, a função de verossimilhança para os dados completos será $\ell_c(\boldsymbol{\theta}|\mathbf{y}_c) = \sum_{i=1}^m \ell_i(\boldsymbol{\theta}|\mathbf{y}_c)$, em que

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

e C é uma constante independente do vetor de parâmetros $\boldsymbol{\theta}$. Dada a estimativa atual $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}^{(k)}$, o passo E se resume ao cálculo da esperança condicional da função log-verossimilhança completa dada por

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

em que

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

e

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

com

$$\begin{aligned}
\widehat{a_i}^{(k)} &= tr\left(\widehat{u_i\mathbf{y}_i\mathbf{y}_i^\top}^{(k)}\Psi^{-1(k)} - 2\widehat{u_i\mathbf{y}_i\mathbf{z}_i^\top}^{(k)}\boldsymbol{\Lambda}^{\top(k)}\boldsymbol{\Psi}^{-1(k)} + \boldsymbol{\Lambda}^{(k)}\widehat{u_i\mathbf{z}_i\mathbf{z}_i^\top}^{(k)}\boldsymbol{\Lambda}^{\top(k)}\boldsymbol{\Psi}^{-1}\right), \\
\widehat{u_i\mathbf{z}_i\mathbf{z}_i^\top}^{(k)} &= E\left[u_i\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)}\left(\widehat{u_i\mathbf{y}_i\mathbf{y}_i}^{\top(k)} - \widehat{u_i\mathbf{y}_i^{(k)}}\boldsymbol{\beta}^{\top(k)}\mathbf{X}_i^\top\right)\boldsymbol{\Psi}^{-1(k)}\boldsymbol{\Lambda}^{\top(k)}\boldsymbol{\Delta}^{(k)} \\
&+ \boldsymbol{\Delta}^{(k)}\boldsymbol{\Lambda}^{\top(k)}\boldsymbol{\Psi}^{-1(k)}\left(-\mathbf{X}_i\boldsymbol{\beta}^{(k)}\widehat{u_i\mathbf{y}_i}^{(k)} + \mathbf{X}_i\boldsymbol{\beta}^{(k)}\boldsymbol{\beta}^{\top(k)}\mathbf{X}_i^\top\right)\boldsymbol{\Psi}^{-1(k)}\boldsymbol{\Lambda}^{\top(k)}\boldsymbol{\Delta}^{(k)}, \\
\boldsymbol{\Delta}^{(k)} &= \left(\boldsymbol{\Omega}^{-1(k)} + \boldsymbol{\Lambda}^{\top(k)}\boldsymbol{\Psi}^{-1(k)}\boldsymbol{\Lambda}^{(k)}\right), \\
\widehat{u_i\mathbf{z}_i}^{(k)} &= E\left[u_i\mathbf{z}_i|\mathbf{V}_i, \mathbf{C}_i, \boldsymbol{\theta}^{(k)}\right] = \boldsymbol{\Delta}^{(k)}\boldsymbol{\Lambda}^{\top(k)}\boldsymbol{\Psi}^{-1(k)}\left(\widehat{u_i\mathbf{y}_i}^{(k)} - \mathbf{X}_i\boldsymbol{\beta}^{(k)}\right), \\
\widehat{u_i\mathbf{y}_i\mathbf{z}_i^\top}^{(k)} &= E\left[u_i\mathbf{y}_i\mathbf{z}_i^\top|\mathbf{V}_i, \mathbf{C}_i, \boldsymbol{\theta}^{(k)}\right] = (\widehat{u_i\mathbf{y}_i\mathbf{y}_i}^{\top(k)} - \widehat{u_i\mathbf{y}_i}^{(k)}\boldsymbol{\beta}^{\top(k)}\mathbf{X}_i^\top)\boldsymbol{\Psi}^{-1(k)}\boldsymbol{\Lambda}^{(k)}\boldsymbol{\Delta}^{(k)}.
\end{aligned}$$

O passo de maximização condicional(CM) de $Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k)})$ é dado por:

$$\boldsymbol{\beta}^{(k+1)} = (\sum_{i=1}^{n} \widehat{u}_i^{(k)} \mathbf{X}_i^\top \mathbf{X}_i)^{-1} \sum_{i=1}^{n} \mathbf{X}_i^\top (\widehat{u_i \mathbf{y}_i}^{(k)} - \boldsymbol{\Lambda}^{(k)} \widehat{u_i \mathbf{z}_i}^{(k)}) \tag{3.3.3}$$

$$\boldsymbol{\Lambda}_j^{(k+1)} = (\sum_{i=1}^{n} \widehat{u_i \mathbf{z}_i \mathbf{z}_i^\top}^{(k)})^{-1} \sum_{i=1}^{n} \left[(\widehat{u_i \mathbf{y}_i \mathbf{z}_i^\top})_j^{\top(k)} - \widehat{u_i \mathbf{z}_i}^{(k)} (\mathbf{X}_i \boldsymbol{\beta}^{(k)})_j \right] \tag{3.3.4}$$

$$\boldsymbol{\Psi}^{(k+1)} = Diag\left(\frac{\sum_{i=1}^{n} (\mathbf{A}_i^{(k)} + \mathbf{A}_i^{(k)\top})}{2n} \right), \tag{3.3.5}$$

em que $\boldsymbol{\Lambda}_j^\top$ é a j-ésima linha de $\boldsymbol{\Lambda}$ para $j = 1, \ldots, p$ e $\mathbf{A}_i^{(k)} = \widehat{u_i \mathbf{y}_i \mathbf{y}_i^\top}^{(k)} - 2\widehat{u_i \mathbf{y}_i}^{(k)} \boldsymbol{\beta}^{\top(k)} \mathbf{X}_i^\top - 2\widehat{u_i \mathbf{y}_i \mathbf{z}_i^\top}^{(k)} \boldsymbol{\Lambda}^{\top(k)} + 2\boldsymbol{\Lambda}^{(k)} \widehat{u_i \mathbf{z}_i}^{(k)} \boldsymbol{\beta}^{\top(k)} \mathbf{X}_i^\top + \mathbf{X}_i \boldsymbol{\beta}^{(k)} \boldsymbol{\beta}^{\top(k)} \mathbf{X}_i^\top + \boldsymbol{\Lambda}^{(k)} \widehat{u \mathbf{z}_i \mathbf{z}_i^\top}^{(k)} \boldsymbol{\Lambda}^{\top(k)}$. O passos E e M do algoritmo são processados de forma iterativa até que algum critério de parada seja estabelecido. Como já descrito, medidas de distância que envolvam duas avaliações sucessivas da função de verossimilhança são comumente utilizadas, tais como $|\ell(\widehat{\boldsymbol{\theta}}^{(k+1)}) - \ell(\widehat{\boldsymbol{\theta}}^{(k)})|$ ou $|\ell(\widehat{\boldsymbol{\theta}}^{(k+1)})/\ell(\widehat{\boldsymbol{\theta}}^{(k)}) - 1|$.

Das Equações (3.3.3) a (3.3.5), verifica-se que o passo E está associado basicamente ao cálculo de $\widehat{u_i \mathbf{y}_i}^{(k)}$, $\widehat{u_i \mathbf{y}_i \mathbf{y}_i^\top}^{(k)}$ e $\widehat{u}_i^{(k)}$. Esses valores esperados podem ser obtidos de forma fechada, utilizando as Proposições 1-3 apresentadas no Apêndice C:

- Da Proposição 1 (Apêndice C), se o i-ésimo indivíduo tiver somente respostas censuradas:

$$\widehat{u_i \mathbf{y}_i \mathbf{y}_i^\top} = E\left[u_i \mathbf{y}_i \mathbf{y}_i^\top | \mathbf{V}_i, \mathbf{C}_i, \widehat{\boldsymbol{\theta}} \right] = \frac{T_p(\mathbf{V}_i | \widehat{\boldsymbol{\mu}_i}, \widehat{\boldsymbol{\Sigma}}_i^*, \nu + 2)}{T_p(\mathbf{V}_i | \widehat{\boldsymbol{\mu}}_i, \widehat{\boldsymbol{\Sigma}}_i^*, \nu)} E\left[\mathbf{W}_i \mathbf{W}_i^\top \right],$$

$$\widehat{u_i \mathbf{y}_i} = E\left[u_i \mathbf{y}_i | \mathbf{V}_i, \mathbf{C}_i, \widehat{\boldsymbol{\theta}} \right] = \frac{T_p(\mathbf{V}_i | \widehat{\boldsymbol{\mu}}_i, \widehat{\boldsymbol{\Sigma}}_i^*, \nu + 2)}{T_p(\mathbf{V}_i | \widehat{\boldsymbol{\mu}}_i, \widehat{\boldsymbol{\Sigma}}_i^*, \nu)} E\left[\mathbf{W}_i \right],$$

$$\widehat{u_i} = E\left[u_i | \mathbf{V}_i, \mathbf{C}_i, \widehat{\boldsymbol{\theta}} \right] = \frac{T_p(\mathbf{V}_i | \widehat{\boldsymbol{\mu}}_i, \widehat{\boldsymbol{\Sigma}}_i^*, \nu + 2)}{T_p(\mathbf{V}_i | \widehat{\boldsymbol{\mu}}_i, \widehat{\boldsymbol{\Sigma}}_i^*, \nu)},$$

  em que $\mathbf{W}_i \sim Tt_p(\widehat{\boldsymbol{\mu}_i}, \widehat{\boldsymbol{\Sigma}}_i^*, \nu + 2r, \mathbb{A}_i)$, $\widehat{\boldsymbol{\mu}}_i = \mathbf{X}_i \widehat{\boldsymbol{\beta}}$, $\widehat{\boldsymbol{\Sigma}}_i^* = \dfrac{\nu}{\nu + 2} \widehat{\boldsymbol{\Sigma}}_i$, $\widehat{\boldsymbol{\Sigma}}_i = \widehat{\boldsymbol{\Psi}} + \widehat{\boldsymbol{\Lambda}} \widehat{\boldsymbol{\Omega}} \widehat{\boldsymbol{\Lambda}}^\top$ e $\mathbb{A} = \{\mathbf{W}_i = (w_1, \ldots, w_p)^\top | w_1 \leq V_{i1}, \ldots, w_p \leq V_{ip}\}$.

- Se o i-ésimo indivíduo tiver todas as respostas observadas (isto é, não censuradas), então:

$$\widehat{u \mathbf{y}_i \mathbf{y}_i^\top} = \frac{\nu + p}{\nu + \delta_i} \mathbf{y}_i \mathbf{y}_i^\top, \ \widehat{u \mathbf{y}_i} = \frac{\nu + p}{\nu + \delta_i} \mathbf{y}_i \ \text{e} \ \widehat{u}_i = \frac{\nu + p}{\nu + \delta_i},$$

  em que $\delta_i = (\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta})^\top \boldsymbol{\Sigma}_i^{-1} (\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta})$.

- Se o i-ésimo indivíduo tiver respostas censuradas e não censuradas, dada a Proposição 2 do Apêndice C e o fato de que $\mathbf{y}_i | \mathbf{V}_i, \mathbf{C}_i$, $\mathbf{y}_i | \mathbf{V}_i, \mathbf{C}_i, \mathbf{y}_i^o$ e $\mathbf{y}_i^c | \mathbf{V}_i, \mathbf{C}_i, \mathbf{y}_i^o$ são equivalentes, tem-se:

$$\widehat{u \mathbf{y}_i \mathbf{y}_i^\top} = E\left[u_i \mathbf{y}_i \mathbf{y}_i^\top | \mathbf{y}_i^o, \mathbf{V}_i, \mathbf{C}_i, \hat{\boldsymbol{\theta}} \right] = \begin{pmatrix} \widehat{u}_i \mathbf{y}_i^o \mathbf{y}_i^{o\top} & \widehat{u}_i \mathbf{y}_i^o \widehat{\mathbf{w}_i^c}^\top \\ \widehat{u}_i \widehat{\mathbf{w}_i^c} \mathbf{y}_i^{o\top} & \widehat{u}_i \widehat{\mathbf{w}_i^c \mathbf{w}_i^{c\top}} \end{pmatrix},$$

$$\widehat{u \mathbf{y}_i} = E\left[u_i \mathbf{y}_i | \mathbf{y}_i^o, \mathbf{V}_i, \mathbf{C}_i, \hat{\boldsymbol{\theta}} \right] = vec(\widehat{u}_i \mathbf{y}_i^o, \widehat{\mathbf{w}_i^c}),$$

$$\widehat{u}_i = E\left[u_i | \mathbf{y}_i^o, \mathbf{V}_i, \mathbf{C}_i, \hat{\boldsymbol{\theta}} \right] = \left(\frac{\nu + p_i^o}{\nu + \delta_i^o} \right) \frac{T_{p_i^c}(\mathbf{V}_i | \boldsymbol{\mu}_i^{co}, \tilde{\mathbf{S}}_i^{co}, \nu + p_i^o + 2)}{T_{p_i^c}(\mathbf{V}_i | \boldsymbol{\mu}_i^{co}, \boldsymbol{\Sigma}_i^{co}, \nu + p_i^o)},$$

em que $\tilde{\mathbf{S}}_i^{co} = \left(\dfrac{\nu+\delta_i^o}{\nu+2+p_i^o}\right)\boldsymbol{\Sigma}_i^{cc.o}$, $\widehat{\mathbf{w}_i^c} = E[\mathbf{W}_i]$ e $\widehat{\mathbf{w}_i^c\mathbf{w}_i^{c\top}} = E[\mathbf{W}_i\mathbf{W}_i^\top]$, com $\mathbf{W}_i \sim Tt_{p_i^c}(\boldsymbol{\mu}_i^{co}, \tilde{\mathbf{S}}_i^{co}, \nu+p_i^o+2, \mathbb{A}_i^c)$ e $\boldsymbol{\mu}_i^{co}$,$\boldsymbol{\Sigma}_i^{cc.o}$ e $\mathbf{S}_i^{co}$ como definido anteriormente.

3.3.2 Estimação da variância dos efeitos fixos

Medidas de erro-padrão para estimativas por máxima verossimilhança podem ser aproximadas pela inversa da matriz de informação observada, mas, geralmente, esta não possui uma expressão fechada. Vaida et al. (2007, Sec. 2) sugeriram uma adaptação da fórmula de Louis (Louis, 1982) para se obter uma matriz de variância ajustada para $\widehat{\boldsymbol{\beta}}$ na presença de dados censurados. Esse procedimento será utilizado aqui e também foi adotado por Vaida and Liu (2009, Sec.2) e Matos et al. (2013b, Sec.3). A estimativa da matriz de variância-covariância de $\widehat{\boldsymbol{\beta}}$ é dada pela matriz

$$\mathbf{J}_{\boldsymbol{\beta}\boldsymbol{\beta}} = Var(\widehat{\boldsymbol{\beta}}) =$$
$$\left(\sum_{i=1}^m \left(\frac{\nu+p}{\nu+p+2}\right)\mathbf{X}_i^\top\boldsymbol{\Sigma}_i^{-1}\mathbf{X}_i - \mathbf{X}_i^\top\boldsymbol{\Sigma}_i^{-1}Var\left[\left(\frac{\nu+p}{\nu+\delta_i}\right)(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta})|\mathbf{V}_i,\mathbf{C}_i\right]\boldsymbol{\Sigma}_i^{-1}\mathbf{X}_i\right)^{-1},$$

em que $\mathbf{y}_i \sim Tt_p(\mathbf{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma}, \nu; \mathbb{A})$. Dado que $Var\left[\left(\frac{\nu+p}{\nu+\delta_i}\right)(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta})|\mathbf{V}_i,\mathbf{C}_i\right]$ depende de $\widehat{u_i\mathbf{y}_i^2} = E\left[\left(\frac{\nu+p}{\nu+\delta_i}\right)^2\mathbf{y}_i\mathbf{y}_i^\top|\mathbf{V}_i,\mathbf{C}_i,\hat{\boldsymbol{\theta}}\right]$, $\widehat{u_i\mathbf{y}_i^1} = E\left[\left(\frac{\nu+p}{\nu+\delta_i}\right)^2\mathbf{y}_i|\mathbf{V}_i,\mathbf{C}_i,\hat{\boldsymbol{\theta}}\right]$ e $\widehat{u_i\mathbf{y}_i^0} = E\left[\left(\frac{\nu+p}{\nu+\delta_i}\right)^2|\mathbf{V}_i,\mathbf{C}_i,\hat{\boldsymbol{\theta}}\right]$, após algumas manipulações algébricas (veja Apêndice C), tem-se

- Se o indivíduo i tem somente respostas censuradas, então pela Proposição 1 (Apêndice C)

$$\widehat{u_i\mathbf{y}_i^2} = c_p(\nu,2)\frac{T_p(\mathbf{V}_i|\widehat{\boldsymbol{\mu}}_i,\widehat{\boldsymbol{\Sigma}}_i^{**},\nu+4)}{T_p(\mathbf{V}_i|\widehat{\boldsymbol{\mu}}_i,\widehat{\boldsymbol{\Sigma}}_i^{*},\nu)}E\left[\mathbf{W}_i\mathbf{W}_i^\top\right],$$
$$\widehat{u_i\mathbf{y}_i^1} = c_p(\nu,2)\frac{T_p(\mathbf{V}_i|\widehat{\boldsymbol{\mu}}_i,\widehat{\boldsymbol{\Sigma}}_i^{**},\nu+4)}{T_p(\mathbf{V}_i|\widehat{\boldsymbol{\mu}}_i,\widehat{\boldsymbol{\Sigma}}_i^{*},\nu)}E\left[\mathbf{W}_i\right],$$
$$\widehat{u_i\mathbf{y}_i^0} = c_p(\nu,2)\frac{T_p(\mathbf{V}_i|\widehat{\boldsymbol{\mu}}_i,\widehat{\boldsymbol{\Sigma}}_i^{**},\nu+4)}{T_p(\mathbf{V}_i|\widehat{\boldsymbol{\mu}}_i,\widehat{\boldsymbol{\Sigma}}_i^{*},\nu)},$$

  em que $\mathbf{W}_i \sim Tt_p(\widehat{\boldsymbol{\mu}}_i, \widehat{\boldsymbol{\Sigma}}_i^*, \nu+4, \mathbb{A})$, $\widehat{\boldsymbol{\Sigma}}_i^{**} = \dfrac{\nu}{\nu+4}\widehat{\boldsymbol{\Sigma}}_i$ e

$$c_p(\nu,2) = \left(\frac{\nu+p}{\nu}\right)^2\left[\frac{\Gamma\left(\frac{\nu+p}{2}\right)\Gamma\left(\frac{\nu+4}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\Gamma\left(\frac{\nu+p+4}{2}\right)}\right].$$

- Se o i-ésimo indivíduo tem somente componentes não censurados, então

$$\widehat{u_i\mathbf{y}_i^2} = \left(\frac{\nu+p}{\nu+\delta_i}\right)^2\mathbf{y}_i\mathbf{y}_i^\top,\quad \widehat{u_i\mathbf{y}_i^1} = \left(\frac{\nu+p}{\nu+\delta_i}\right)^2\mathbf{y}_i \text{ e } \widehat{u_i\mathbf{y}_i^0} = \left(\frac{\nu+p}{\nu+\delta_i}\right)^2,$$

  em que $\delta_i = (\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})^\top\boldsymbol{\Sigma}_i^{-1}(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta})$.

- Se o i-ésimo indivíduo tiver componentes censurados e não censurados, então, pela Proposição 2 do Apêndice C, tem-se:

$$
\begin{aligned}
\widehat{u_i\mathbf{y}_i^2} &= \begin{pmatrix} \widehat{u_i\mathbf{y}_i^0}\,\mathbf{y}_i^o\mathbf{y}_i^{o\top} & \widehat{u\mathbf{y}_i^0}\,\mathbf{y}_i^o\,\widehat{\mathbf{w}_i^c}^\top \\ \widehat{u_i\mathbf{y}_i^0}\,\widehat{\mathbf{w}_i^c}\,\mathbf{y}_i^{o\top} & \widehat{u\mathbf{y}_i^0}\,\widehat{\mathbf{w}_i^c\,\mathbf{w}_i^{c\top}} \end{pmatrix}, \\
\widehat{u\mathbf{y}_i^1} &= vec(\widehat{u_i\mathbf{y}_i^0}\,\mathbf{y}_i^o,\, \widehat{\mathbf{w}_i^c}), \\
\widehat{u\mathbf{y}_i^0} &= \left(\frac{d_p}{(\nu+\delta_i^o)^2}\right)\frac{T_{p_i^c}(\mathbf{V}_i|\boldsymbol{\mu}_i^{co}, \tilde{\mathbf{S}}_i^{co}, \nu+p_i^o+4)}{T_{p_i^c}(\mathbf{V}_i|\boldsymbol{\mu}_i^{co}, \boldsymbol{\Sigma}_i^{co}, \nu+p_i^o)},
\end{aligned}
$$

em que $d_p = (\nu+p)^2\left(\frac{\Gamma((p+\nu)/2)\Gamma((p_1+\nu+4)/2)}{\Gamma((p+\nu)/2)\Gamma((p+\nu+4)/2)}\right)$, $\tilde{\mathbf{S}}^{co} = \left(\frac{\nu+\delta_i^o}{\nu+2+p_i^o}\right)\boldsymbol{\Sigma}_i^{cc.o}$, $\widehat{\mathbf{w}_i^c} = E[\mathbf{W}_i]$ e $\widehat{\mathbf{w}_i^c\mathbf{w}_i^{c\top}} = E[\mathbf{W}_i\mathbf{W}_i^\top]$, com $\mathbf{W}_i \sim Tt_{p_i^c}(\boldsymbol{\mu}_i^{co}, \tilde{\mathbf{S}}_i^{co}, \nu+p_i^o+2, \mathbb{A}_i^c)$ e $\boldsymbol{\mu}_i^{co}$,$\boldsymbol{\Sigma}_i^{cc.o}$ e $\mathbf{S}_i^{co}$ como definido anteriormente.

Intervalos de confiança assintóticos e testes de hipóteses para os efeitos fixos podem ser obtidos assumindo que as estimativas de máxima verossimilhança $\widehat{\boldsymbol{\beta}}$ seguem aproximadamente uma distribuição $N_p(\boldsymbol{\beta}, \mathbf{J}_{\boldsymbol{\beta}\boldsymbol{\beta}}^{-1})$. Na prática, $\mathbf{J}_{\boldsymbol{\beta}\boldsymbol{\beta}}$ é comumente desconhecida e precisa ser substituída por sua estimativa de máxima verossimilhança $\mathbf{J}_{\widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{\beta}}}$.

3.3.3 Estimação dos fatores latentes

Da mesma forma que apresentado no Capítulo 2, para estimação dos fatores latentes associados ao modelo TCFA foi considerada uma abordagem condicional tal como descrito por (Matos et al., 2013b). Dadas as estimativas de $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \boldsymbol{\Lambda}, \Psi)^\top$ e considerando ν conhecido, a média de $\mathbf{z}_i$ condicionada a $\mathbf{V}_i$ e $\mathbf{C}_i$ é:

$$
\begin{aligned}
\widehat{\mathbf{z}}_i(\boldsymbol{\theta}) &= E[\mathbf{z}_i|\mathbf{V}_i, \mathbf{C}_i] = E\left[E[E(\mathbf{z}_i|u_i)|\mathbf{y}_i, u_i]|\mathbf{V}_i, \mathbf{C}_i\right] = E\{\boldsymbol{\Delta}\boldsymbol{\Lambda}^\top\boldsymbol{\Psi}^{-1}\left(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}\right)|\mathbf{V}_i, \mathbf{C}_i\} \\
&= \boldsymbol{\Delta}\boldsymbol{\Lambda}^\top\boldsymbol{\Psi}^{-1}\left(\widehat{\mathbf{y}_i} - \mathbf{X}_i\boldsymbol{\beta}\right),
\end{aligned}
$$

em que $\boldsymbol{\Delta} = \left(\boldsymbol{\Omega}^{-1} + \boldsymbol{\Lambda}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}\right)$ e $\widehat{\mathbf{y}_i} = E\{\mathbf{y}_i|\mathbf{V}_i, \mathbf{C}_i\}$ é o primeiro momento de uma distribuição multivariada t de Student truncada $Tt_p(\mathbf{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma}_i, \nu; \mathbb{A}_i)$. Na prática, estimativas dos fatores latentes, $\mathbf{z}_i$, são obtidas substituindo as estimativas de máxima verossimilhança $\widehat{\boldsymbol{\theta}}$, implicando em $\widehat{\mathbf{z}_i} = \widehat{\mathbf{z}}_i(\widehat{\boldsymbol{\theta}})$.

A matriz de variância-covariância condicional de $\mathbf{z}_i$ dados $\mathbf{V}_i$ e $\mathbf{C}_i$ é

$$
\begin{aligned}
Var[\mathbf{z}_i|\mathbf{V}_i, \mathbf{C}_i] &= E[\mathbf{z}_i\mathbf{z}_i^\top|\mathbf{V}_i, \mathbf{C}_i] - \widehat{\mathbf{z}}_i(\boldsymbol{\theta})\widehat{\mathbf{z}}_i(\boldsymbol{\theta})^\top \\
&= \boldsymbol{\Delta}E\left[\left(\frac{\nu+p}{\nu+\delta_i}\right)^{-1}|\mathbf{V}_i, \mathbf{C}_i\right] + \boldsymbol{\Delta}\boldsymbol{\Lambda}^\top\boldsymbol{\Psi}^{-1}Var\left[\left(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}\right)|\mathbf{V}_i, \mathbf{C}_i\right]\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Delta}^\top.
\end{aligned}
$$

Esses valores esperados são facilmente obtidos por meio do passo E do algoritmo ECM proposto.

3.3.4 Aplicações

Nesta seção será apresentado um resumo do estudo de simulação e da análise de dados reais aplicados para avaliação dos métodos desenvolvidos.

Estudo de simulação

O estudo de simulação desenvolvido aqui se assemelha à configuração do estudo de simulação apresentado no Capítulo 2. O objetivo principal é avaliar o impacto das inferências sobre os parâmetros do modelo quando a suposição de normalidade é violada. Para tanto, três cenários foram definidos:

Cenário 1 (TCFA): Os dados e os fatores latentes foram gerados de uma distribuição Normal;

Cenário 2 (t-TCFA): Os dados e os fatores latentes foram gerados de uma distribuição t de Student com 4 graus de liberdade;

Cenário 3 (Normal contaminada): Os dados e os fatores latentes foram gerados de uma distribuição Normal Contaminada com parâmetros $\nu = (\nu_1, \nu_2)^\top = (0.1, 0.1)^\top$.

Para todos os cenários, três tamanhos de amostra foram considerados, $n = 50$, 150 e 300. Os modelos Gaussiano TCFA (Seção 3.2) e o t-TCFA com $\nu = 4$ (Seção 3.3) foram ajustados para cada cenário. Para os 9 diferentes estudos, 100 réplicas dos dados foram geradas considerando a seguinte estrutura:

$$\begin{aligned}
\mathbf{Y}_i|\mathbf{Z}_i, U_i = u_i &\sim N_p(\boldsymbol{\mu}_i, u_i^{-1}\boldsymbol{\Psi}),\\
\mathbf{Z}_i|U_i = u_i &\sim N_q(\mathbf{0}, u_i^{-1}\boldsymbol{\Omega}),\\
U_i &\sim F(\nu),
\end{aligned}$$

em que $\boldsymbol{\mu}_i = \mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\Lambda}\mathbf{z}_i$, com $p = 5$, $k = 3$, $q = 2$ e $\boldsymbol{\beta}^\top = (3.5, 2.5, -1.5)$,

$$\boldsymbol{\Lambda}^\top = \begin{pmatrix} -0.6 & -0.6 & -0.6 & 0 & 0 \\ 0 & 0 & 0 & 0.5 & 0.5 \end{pmatrix},$$

$\boldsymbol{\Psi} = diag(0.3, 0.4, 0.6, 0.2, 0.7)$, e $U_i \equiv 1$ para todo i para o cenário 1 (distribuição Normal), $U_i \sim \text{Gamma}(2, 2)$ para todo i do cenário 2 (distribuição t de Student) e $U_i = \begin{cases} \nu_2 & \text{com prob} \quad \nu_1 \\ 1 & \text{com prob} \quad 1 - \nu_1 \end{cases}$ para todo i do cenário 3 (distribuição Normal contaminada).

Como no estudo anterior, os elementos da matriz $\boldsymbol{\Omega}$ foram fixados em $0.6I + 0.4J$, em que I é uma matriz identidade de dimensão (2×2) e J é uma matriz de uns de dimensão (2×2). A matriz de delineamento X_i foi gerada pelo produto *Kronecker* de uma matriz de uns com dimensão $(1 \times p)$ e um vetor transposto cujos elementos eram x_{ij}, $j = 1, 2, 3$. A covariável x_{i1} foi fixada em 1 para todo $i = 1, \ldots, n$ representando o intercepto; x_{i2} foi gerada de uma distribuição Uniforme discreta com pontos 6, 7, 8, 9, 10 e probabilidades 0.05, 0.4, 0.45, 0.05, 0.05, imitando a variável idade presente na análise dos dados EGRA. Para estabilidade computacional, x_{i2} foi centralizada em 8. A última covariável x_{i3}, por sua vez, foi gerada de uma distribuição de Bernoulli com parâmetro $p = 0.5$, imitando a variável sexo da base de dados reais. Todos os valores $\mathbf{Y}$ menores ou iguais a zero foram considerados como casos censurados.

Cenário 1: Por meio da Tabela 3.3, verifica-se que o modelo sob normalidade (modelo TCFA) obteve melhor desempenho comparado ao t-TCFA quando os dados foram gerados da distribuição Normal. Os critérios de comparação de modelos AIC e BIC apontaram essa preferência, contudo, as estimativas obtidas pelo modelo t-TCFA não destoaram muito de seus valores verdadeiros. Verifica-se também que as medidas de erros-padrão menores são e, consequentemente, as estimativas de ambos modelos são mais precisa à medida que o tamanho de amostra cresce.

Cenário 2: Quando os dados foram gerados da distribuição t de Student, o modelo t-TCFA teve melhor desempenho segundo os critérios de seleção de modelos AIC e BIC (Tabela 3.4). O modelo Gaussiano, por sua vez, superestimou as variâncias específicas ($\boldsymbol{\Psi}$) do modelo para qualquer um dos tamanhos amostrais avaliados. Esse fato fornece indícios da falta de robustez do modelo TCFA. Com relação às medidas de erro-padrão, verificou-se que os erros associados às estimativas obtidas por ambas abordagem diminuíram à medida que se aumentou o tamanho da amostra, tal como observado no cenário 1.

Cenário 3: Nesse cenário, verificou-se o desempenho dos modelos TCFA e t-TCFA quando os dados foram gerados de uma distribuição de caudas mais pesadas que elas. Por meio da Tabela 3.5, os critérios AIC e BIC claramente apontaram a indicação do modelo t-TCFA como o melhor modelo. Mais uma vez, o modelo TCFA apresentou estimativas altas para os elementos da matriz $\boldsymbol{\Psi}$. Observando as medidas de vício e erro quadrático médio associados às estimativas obtidas nesse cenário, é evidente que o modelo t-TCFA estimou melhor os parâmetros do modelo e apresenta-se como uma ótima alternativa para análise robusta de dados que não advêm da distribuição Gaussiana usual.

Tabela 3.3: **Cenário 1.** MC mean e MC Sd (em parênteses) são, respectivamente, as estimativas médias via Monte Carlo e os desvios-padrão ao se ajustar os modelos TCFA e t-TCFA. Os valores em parênteses são as estimativas médias do erro-padrão aproximado obtidas pelo método baseado na função de informação. MC AIC e MC BIC são os valores médios das respectivos critérios de comparação de modelos.

		TCFA		t-TCFA	
Tamanho da amostra	Parâmetros	Mc mean	MC Sd	Mc mean	MC Sd
n=50	β_1	3.495 (0.097)	0.102	3.48 (0.088)	0.104
	β_2	2.507 (0.067)	0.062	2.494 (0.063)	0.062
	β_3	-1.495 (0.135)	0.141	-1.476 (0.124)	0.138
	λ_{11}	-0.546	0.104	-0.512	0.106
	λ_{21}	-0.562	0.127	-0.531	0.128
	λ_{31}	-0.573	0.165	-0.548	0.159
	λ_{42}	0.290	0.147	0.326	0.102
	λ_{52}	0.437	0.216	0.537	0.144
	ψ_{11}	0.317	0.090	0.274	0.082
	ψ_{22}	0.369	0.099	0.319	0.088
	ψ_{33}	0.581	0.167	0.495	0.149
	ψ_{44}	0.335	0.106	0.260	0.077
	ψ_{55}	0.688	0.170	0.493	0.105
	MC AIC	530.745		552.484	
	MC BIC	555.601		577.340	
n=150	β_1	3.493 (0.053)	0.049	3.483 (0.048)	0.050
	β_2	2.500 (0.043)	0.043	2.492 (0.040)	0.046
	β_3	-1.484 (0.079)	0.076	-1.476 (0.070)	0.076
	λ_{11}	-0.563	0.064	-0.531	0.062
	λ_{21}	-0.579	0.081	-0.551	0.080
	λ_{31}	-0.590	0.089	-0.566	0.084
	λ_{42}	0.332	0.088	0.340	0.061
	λ_{52}	0.496	0.118	0.555	0.062
	ψ_{11}	0.312	0.054	0.270	0.050
	ψ_{22}	0.382	0.061	0.328	0.052
	ψ_{33}	0.589	0.089	0.503	0.080
	ψ_{44}	0.312	0.061	0.248	0.044
	ψ_{55}	0.667	0.110	0.494	0.071
	MC AIC	1620.169		1691.247	
	MC BIC	1659.307		1730.386	
n=300	β_1	3.499 (0.039)	0.038	3.488 (0.035)	0.037
	β_2	2.501 (0.032)	0.033	2.491 (0.030)	0.034
	β_3	-1.500 (0.056)	0.050	-1.488 (0.050)	0.049
	λ_{11}	-0.570	0.047	-0.540	0.045
	λ_{21}	-0.580	0.057	-0.554	0.057
	λ_{31}	-0.587	0.062	-0.566	0.060
	λ_{42}	0.339	0.046	0.343	0.034
	λ_{52}	0.511	0.068	0.558	0.044
	ψ_{11}	0.313	0.039	0.274	0.034
	ψ_{22}	0.390	0.051	0.337	0.044
	ψ_{33}	0.589	0.065	0.503	0.059
	ψ_{44}	0.314	0.037	0.254	0.027
	ψ_{55}	0.676	0.068	0.511	0.047
	MC AIC	3228.259		3378.058	
	MC BIC	3276.408		3426.207	

Tabela 3.4: **Cenário 2.** MC mean e MC Sd (em parênteses) são, respectivamente, as estimativas médias via Monte Carlo e os desvios-padrão ao se ajustar os modelos TCFA e t-TCFA. Os valores em parênteses são as estimativas médias do erro-padrão aproximado obtidas pelo método baseado na função de informação. MC AIC e MC BIC são os valores médios das respectivos critérios de comparação de modelos.

		TCFA		t-TCFA	
Tamanho da amostra	Parâmetros	Mc mean	MC Sd	Mc mean	MC Sd
n=50	β_1	3.500 (0.133)	0.136	3.481 (0.102)	0.106
	β_2	2.514 (0.091)	0.096	2.480 (0.072)	0.073
	β_3	-1.523 (0.185)	0.190	-1.505 (0.143)	0.149
	λ_{11}	-0.762	0.268	-0.610	0.134
	λ_{21}	-0.786	0.275	-0.621	0.158
	λ_{31}	-0.788	0.332	-0.628	0.196
	λ_{42}	0.416	0.264	0.391	0.143
	λ_{52}	0.607	0.357	0.638	0.170
	ψ_{11}	0.574	0.262	0.358	0.112
	ψ_{22}	0.661	0.300	0.433	0.151
	ψ_{33}	1.037	0.336	0.694	0.209
	ψ_{44}	0.659	0.326	0.339	0.105
	ψ_{55}	1.315	0.685	0.661	0.199
	MC AIC	648.072		639.362	
	MC BIC	672.929		664.218	
n=150	β_1	3.511 (0.073)	0.070	3.485 (0.055)	0.060
	β_2	2.507 (0.059)	0.059	2.484 (0.046)	0.046
	β_3	-1.522 (0.109)	0.100	-1.497 (0.082)	0.082
	λ_{11}	-0.780	0.155	-0.619	0.067
	λ_{21}	-0.821	0.179	-0.647	0.088
	λ_{31}	-0.823	0.225	-0.645	0.115
	λ_{42}	0.484	0.156	0.408	0.078
	λ_{52}	0.731	0.228	0.666	0.088
	ψ_{11}	0.590	0.235	0.349	0.056
	ψ_{22}	0.719	0.219	0.442	0.086
	ψ_{33}	1.024	0.218	0.644	0.105
	ψ_{44}	0.586	0.149	0.324	0.049
	ψ_{55}	1.283	0.319	0.664	0.107
	MC AIC	2004.130		1950.676	
	MC BIC	2043.269		1989.815	
n=300	β_1	3.509 (0.053)	0.054	3.487 (0.040)	0.043
	β_2	2.499 (0.044)	0.046	2.483 (0.034)	0.034
	β_3	-1.507 (0.077)	0.072	-1.492 (0.057)	0.058
	λ_{11}	-0.785	0.110	-0.615	0.053
	λ_{21}	-0.821	0.112	-0.648	0.060
	λ_{31}	-0.824	0.143	-0.652	0.083
	λ_{42}	0.483	0.117	0.404	0.055
	λ_{52}	0.734	0.173	0.657	0.069
	ψ_{11}	0.583	0.143	0.353	0.044
	ψ_{22}	0.724	0.132	0.446	0.059
	ψ_{33}	1.055	0.203	0.649	0.080
	ψ_{44}	0.584	0.115	0.326	0.040
	ψ_{55}	1.283	0.263	0.663	0.077
	MC AIC	3999.002		3885.252	
	MC BIC	4047.151		3933.401	

Tabela 3.5: **Cenário 3.** MC mean e MC Sd (em parênteses) são, respectivamente, as estimativas médias via Monte Carlo e os desvios-padrão ao se ajustar os modelos TCFA e *t*-TCFA. Os valores em parênteses são as estimativas médias do erro-padrão aproximado obtidas pelo método baseado na função de informação. MC AIC e MC BIC são os valores médios das respectivos critérios de comparação de modelos.

		TCFA		*t*-TCFA	
Tamanho da amostra	Parâmetros	Mc mean	MC Sd	Mc mean	MC Sd
n=50	β_1	3.499 (0.127)	0.126	3.479 (0.098)	0.100
	β_2	2.487 (0.087)	0.096	2.477 (0.070)	0.071
	β_3	-1.500 (0.177)	0.200	-1.465 (0.137)	0.158
	λ_{11}	-0.734	0.266	-0.580	0.128
	λ_{21}	-0.766	0.300	-0.608	0.141
	λ_{31}	-0.742	0.328	-0.597	0.173
	λ_{42}	0.436	0.262	0.383	0.106
	λ_{52}	0.628	0.371	0.601	0.137
	ψ_{11}	0.533	0.244	0.330	0.098
	ψ_{22}	0.624	0.294	0.394	0.108
	ψ_{33}	0.993	0.505	0.594	0.190
	ψ_{44}	0.609	0.274	0.323	0.094
	ψ_{55}	1.255	0.638	0.614	0.186
	MC AIC	638.215		620.283	
	MC BIC	663.071		645.139	
n=150	β_1	3.512 (0.072)	0.068	3.484 (0.053)	0.055
	β_2	2.483 (0.058)	0.061	2.478 (0.044)	0.045
	β_3	-1.509 (0.106)	0.106	-1.481 (0.079)	0.087
	λ_{11}	-0.772	0.150	-0.588	0.076
	λ_{21}	-0.788	0.158	-0.612	0.077
	λ_{31}	-0.773	0.188	-0.628	0.104
	λ_{42}	0.462	0.173	0.391	0.069
	λ_{52}	0.664	0.221	0.617	0.072
	ψ_{11}	0.551	0.170	0.335	0.064
	ψ_{22}	0.667	0.187	0.405	0.077
	ψ_{33}	1.044	0.270	0.609	0.103
	ψ_{44}	0.604	0.152	0.313	0.043
	ψ_{55}	1.212	0.369	0.602	0.095
	MC AIC	1983.951		1903.606	
	MC BIC	2023.089		1942.744	
n=300	β_1	3.505 (0.052)	0.049	3.484 (0.038)	0.039
	β_2	2.492 (0.043)	0.042	2.485 (0.033)	0.032
	β_3	-1.505 (0.075)	0.070	-1.492 (0.055)	0.054
	λ_{11}	-0.762	0.116	-0.590	0.057
	λ_{21}	-0.775	0.112	-0.611	0.058
	λ_{31}	-0.781	0.138	-0.622	0.072
	λ_{42}	0.464	0.119	0.389	0.045
	λ_{52}	0.688	0.165	0.624	0.049
	ψ_{11}	0.568	0.127	0.340	0.042
	ψ_{22}	0.691	0.143	0.410	0.047
	ψ_{33}	1.042	0.204	0.613	0.073
	ψ_{44}	0.577	0.112	0.305	0.033
	ψ_{55}	1.212	0.266	0.610	0.067
	MC AIC	3944.315		3779.822	
	MC BIC	3992.464		3827.971	

Aplicação a dados reais

Nesta parte do trabalho foram utilizados os mesmos dados reais, bem como as mesmas especificações, do estudo apresentado na Seção 3.2. Essa base se refere ao desempenho de 502 estudantes do Peru em quatro subtestes do exame EGRA. O objetivo principal dessa aplicação é de avaliar se o modelo t-TCFA se adequa melhor aos dados em comparação com o modelo TCFA.

Primeiramente, foi realizada uma avaliação do valor dos graus de liberdade mais adequado à análise. Para tanto, o modelo t-TCFA foi ajustado para diferentes valores de ν e, tal como recomendado por Lange et al. (1989), verificou-se que a função log-verossimilhança atingiu seu máximo quando ν foi igual a 5. Dessa forma, esse valor foi fixado para as análises subsequentes.

Tabela 3.6: Dados EGRA. Estimativas MV sob os modelos TCFA e t-TCFA. Em parênteses, as medidas de erro-padrão.

Parâmetros	t-TCFA	TCFA
β_1	0.276 (0.068)	0.251 (0.093)
β_2	0.655 (0.086)	0.765 (0.119)
β_3	0.785 (0.057)	0.931 (0.079)
β_4	-0.111 (0.044)	-0.123 (0.06)
λ_{11}	0.479	0.565
λ_{21}	0.880	0.967
λ_{31}	0.815	0.915
λ_{41}	0.885	0.961
ψ_{11}	0.457	0.788
ψ_{22}	0.114	0.161
ψ_{33}	0.172	0.224
ψ_{44}	0.138	0.206
AIC	4165.902	4231.904
BIC	4216.525	4282.527
CAIc	4228.525	4294.527

Verifica-se pela Tabela 3.6 que, em geral, as estimativas dos parâmetros segundo o modelo t-TCFA foram menores às obtidas pelo modelo TCFA. Sob as duas abordagem, todas as variáveis incluídas no modelo foram significativas. Apesar de baixa carga fatorial da tarefa 1 (habilidade de reconhecer letras do alfabeto), a existência de um fator comum associada a uma habilidade geral (interpretada como "fluência em Língua Espanhola") pode ser também identificada para o modelo t-TCFA. Segundo os critérios de comparação de modelos avaliados, esse modelo se destaca como o preferido.

Pela análise da log-verossimilhança perfilada, observou-se que o algoritmo ECM proposto obteve convergência e pode-se observar que o modelo t-TCFA obteve melhor desempenho em relação ao modelo TCFA.

Capítulo 4

Considerações Finais

Nesta tese discutimos vários aspectos inferenciais envolvendo dois tipos de modelos de variáveis latentes: o modelo linear generalizado misto para respostas binárias (GLMM) e o modelo de análise fatorial Tobit (TCFA) com covariáveis. Os capítulos que compõem este documento foram estruturados como um resumo dos principais pontos abordados nos três artigos desenvolvidos durante essa pesquisa. Para o leitor, no entanto, é fortemente recomendada à leitura dos apêndices, pois estes apresentam de forma mais detalhada os métodos propostos, bem como toda configuração dos estudos de simulação e das aplicações aos bancos de dados reais, além de apresentar as conclusões de cada trabalho.

Várias propostas de pesquisa poderão ser ainda investigadas a partir dos resultados desta tese, entre eles podemos sugerir:

- Extensão para outros tipos de modelos que envolvam variáveis latentes, tais como os modelos da Teoria de Resposta, modelos espaciais, entre outros. Uma prévia do desenvolvimento de métodos robustos para modelagem de dados binários no modelo unidimensional de dois parâmetros da Teoria de Resposta ao Item pode ser vista no Apêndice D.

- Algumas direções envolvendo os modelos GLMM e TCFA podem ser exploradas na sequência desse trabalho, entre elas o utilização de distribuições que além de realizar uma análise robusta à observações discrepantes possa incorporar características como assimetria dos dados. As distribuições de misturas de escala $skew-normal$ (SMSN) têm sido trabalhadas em modelos lineares e não lineares [veja, por exemplo, (Zeller, 2009) and (Lachos et al., 2013a)] e podem contribuir ainda mais no contexto de modelos de variáveis latentes.

- Por fim, outros tópicos de pesquisa ainda poderiam ser desenvolvidos, tais como: a análise de diagnóstico do modelo t-TCFA, extensão para o modelo de análise fatorial exploratória com dados censurados, explorando o problema de identificabilidade desses modelos, melhoria computacional dos algoritmos propostos, bem como a estimação via análise bayesiana para o modelo t-TCFA.

Apêndice A

Generalized linear mixed models for correlated binary data with t-link

Generalized linear mixed models for correlated binary data with t-link

Accepted manuscript - Statistics and Computing

Abstract

A critical issue in modeling binary response data is the choice of the links. We introduce a new link based on the Student's t-distribution (t-link) for correlated binary data. The t-link relates to the common probit-normal link adding one additional parameter which controls the heaviness of the tails of the link. We propose an interesting EM algorithm for computing the maximum likelihood for generalized linear mixed t-link models for correlated binary data. In contrast with recent developments (Tan et al., 2007; Meza et al., 2009), this algorithm uses closed-form expressions at the E-step, as opposed to Monte Carlo simulation. Our proposed algorithm relies on available formulas for the mean and variance of a truncated multivariate t-distribution. To illustrate the new method, a real data set on respiratory infection in children and a simulation study are presented.

Keywords: Correlated binary data, EM-algorithm, Generalized linear mixed models, Truncated multivariate t-distribution

1 Introduction

Generalized linear mixed models (GLMM) (Breslow and Clayton, 1993) are natural extensions of generalized linear models (GLM) when analyzing non-Gaussian data collected from different clusters or from longitudinal studies, in which, population characteristics can be modeled as fixed effects and individual variations as random effects. GLMM apply to either continuous or discrete data. Regarding the latter, a popular class of GLMM is the probit-normal (hereafter probit) model for analyzing binary as well as ordinal data. Although likelihood-based methods for GLMM for Gaussian responses are well developed (Meng and van Dyk, 1998), maximum likelihood (ML) for fitting GLMM for correlated binary responses remains a challenge because of the complexity of the likelihood function. In this context, a great deal of recent attention has focused on the development of efficient methods to maximize the likelihood of GLMM for correlated binary outcomes. For instance, McCulloch (1994) proposed a Monte Carlo EM (MCEM) algorithm with a Gibbs sampler at each E-step. Later, McCulloch (1997) used a Metropolis–Hastings algorithm at each E-step in the MCEM to fit more general models. Tan et al. (2007) proposed a non-iterative importance sampling approach to evaluate the first and the second order moments of a truncated multivariate normal distribution associated with the MCEM algorithm. Meza et al. (2009) proposed to use the stochastic approximation version of EM (SAEM), proposed by Delyon et al. (1999), to obtain the ML estimates. Lee and Nelder (2006) proposed an approximated procedure based on the h-likelihood. However, by their nature, MCEM methods are expensive propositions, due to a combination of Monte Carlo simulation with iterative procedures. In addition, this popular link does not always provide the best fit for a given dataset. In this case, the link could be misspecified, which can yield a substantial bias in the mean response estimates (Czado and Santner, 1992).

One popular way of guarding against the misspecification of links is to embed the probit and logit links, into a more general parametric class of links. Following Liu (2004), in this paper we propose a robust parametric model of GLMM for binary data based on the symmetric t-link, so that the t-GLMM is defined. A full likelihood based approach is carried out, including the implementation of an exact ECM algorithm for maximum likelihood (ML) estimation. Under our proposition, the probit and logit links can be considered as special cases. As in Matos et al. (2013), we show that the E-step reduces to computing the first two moments of certain truncated multivariate t-distributions. The general formulas for these moments were derived by Ho et al. (2012) (eq. 12 and 13). They require the multivariate t cumulative density function (cdf), for which we use the mvtnorm package (Genz et al., 2008) in R (R Core Team, 2013). The likelihood function is easily computed as a by-product of the E-step and is used for monitoring convergence and for model selection, such as, the Akaike information criterion (AIC), the Bayesian information criterion (BIC) and the likelihood ratio test (LRT). To monitor the convergence of the proposed EM, we follow Tan et al. (2007) and directly calculate the log-likelihood values and then plot the difference of the consecutive log-likelihood values against the EM iteration. The numerical results show that the t-link outperforms the probit link in GLMM for correlated binary data under different scenarios.

The rest of this article is organized as follows. In Section 2, we introduce some notation and outline the main results related to the multivariate and truncated t-distributions. The t-GLMM for correlated binary data is formulated in Section 3. The new proposed EM algorithm is developed in Section 4, as well as some inferential results. In Section 5, we present a simulation study and analyze a real data set from a cohort study of Indonesian preschool children for the presence of respiratory infection to illustrate the proposed methods. We conclude with a discussion in Section 6.

2 The multivariate t and truncated t-distribution

A random variable $\mathbf{Y}$ is said to follow a p-variate t-distribution with location vector $\boldsymbol{\mu}$, scale matrix $\boldsymbol{\Sigma}$ and degrees of freedom ν, denoted by $t_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$, if it can be represented by

$$\mathbf{Y} = \boldsymbol{\mu} + U^{-1/2}\mathbf{Z}, \ \ \mathbf{Z} \sim N_p(\mathbf{0}, \boldsymbol{\Sigma}), \ U \sim \text{Gamma}(\nu/2, \nu/2),$$

where $\mathbf{Z}$ and U are independent and Gamma(a, b) stands for a gamma distribution with mean a/b and density denoted by $G(.|a, b)$. We then obtain the probability density function (pdf) of $\mathbf{Y}$, given by

$$t_p(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu) = \frac{\Gamma(\frac{p+\nu}{2})\nu^{-p/2}}{\Gamma(\frac{\nu}{2})\pi^{p/2}}|\boldsymbol{\Sigma}|^{-1/2}\left(1 + \frac{\delta}{\nu}\right)^{-\frac{(p+\nu)}{2}},$$

where $\Gamma(.)$ is the standard gamma function and $\delta = (\mathbf{y} - \boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \boldsymbol{\mu})$ is the Mahalanobis distance. The cdf will be denoted by $T_p(.|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$. If $\nu > 1$, $\boldsymbol{\mu}$ is the mean of $\mathbf{Y}$, and if $\nu > 2$, $\nu(\nu-2)^{-1}\boldsymbol{\Sigma}$ is its covariance matrix. As ν tends to infinity, U converges to one with probability one, and so $\mathbf{Y}$ becomes marginally multivariate normal with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Now, let $Tt_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu; \mathbb{A})$ represent a p-variate truncated t-distribution for $t_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$ lying within a right-truncated hyperplane

$$\mathbb{A} = \{\mathbf{x} = (x_1, \dots, x_p)^\top | x_1 \leq a_1, \dots, x_p \leq a_p\}. \tag{1}$$

Specifically, we say that the p-dimensional vector $\mathbf{X} \sim Tt_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu; \mathbb{A})$ if its density is given by $f(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu; \mathbb{A}) = \dfrac{t_p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)}{T_p(\mathbf{a}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)}\mathbb{I}_{\mathbb{A}}(\mathbf{x})$, where $\mathbf{a} = (a_1, \dots, a_p)^\top$ and $\mathbb{I}_{\mathbb{A}}(\mathbf{x})$ is the indicator function, whose value equals one if $\mathbf{x} \in \mathbb{A}$ and zero otherwise. The following result is related to the moments of the right-truncated multivariate t-distribution and will be useful in the implementation of the EM algorithm in t-GLMM for correlated binary data. The proof is given in the Appendix.

Proposition 2.1. *If* $\mathbf{X} \sim Tt_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu; \mathbb{A})$ *with* $\mathbb{A}$ *as defined in (1), then the kth* *moment of* $\mathbf{X}$, $k = 0, 1, 2$, *is given by*

$$\begin{aligned} E\left[\left(\frac{\nu+p}{\nu+\delta}\right)^r \mathbf{X}^{(k)}\right] &= c_p(\nu, r) \times \\ &\quad \frac{T_p(\mathbf{a}|\boldsymbol{\mu}, \boldsymbol{\Sigma}^*, \nu+2r)}{T_p(\mathbf{a}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)} E_{\mathbf{W}}\left[\mathbf{W}^{(k)}\right], \end{aligned}$$

$$\mathbf{W} \sim Tt_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}^*, \nu+2r; \mathbb{A}),$$

where $c_p(\nu, r) = \left(\frac{\nu+p}{\nu}\right)^r \left(\frac{\Gamma((p+\nu)/2)\Gamma((\nu+2r)/2)}{\Gamma(\nu/2)\Gamma((p+\nu+2r)/2)}\right)$, $\delta = (\mathbf{X}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{X}-\boldsymbol{\mu})$, $\mathbf{a} = (a_1, \ldots, a_p)^\top$, $\boldsymbol{\Sigma}^* = \frac{\nu}{\nu+2r}\boldsymbol{\Sigma}$, $\mathbf{X}^{(0)} = 1$, $\mathbf{X}^{(1)} = \mathbf{X}$, $\mathbf{X}^{(2)} = \mathbf{X}\mathbf{X}^\top$ *and* $\nu + 2r > 0$.

Formulas for $E[\mathbf{W}]$ and $E[\mathbf{W}\mathbf{W}^\top]$, where $\mathbf{W} \sim Tt_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu; \mathbb{A})$, have been recently developed in closed form by Ho et al. (2012), which depend on the multivariate t cdf. The computation uses existing functions for the cumulative t-distribution, for which the *pmvt*() function of the *mvtnorm* library (Genz et al., 2008) from R can be used.

3 The model

3.1 The probit-GLMM for binary data

Let Y_{ij} denote the binary outcome 0 or 1 of the jth measurement and $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{in_i})^\top$ be the collection of responses from subject i, where $i = 1, \ldots, m$ and $j = 1, \ldots, n_i$. The generalized linear mixed probit model (McCulloch, 1994) assumes that given the random effects $\mathbf{b}_i$, the responses $\{Y_{ij}\}_{j=1}^{n_i}$ are conditionally independent with probability

$$\begin{aligned} Pr(Y_{ij} &= 1|\mathbf{b}_i) = \Phi(\mu_{ij}), \;\; \mu_{ij} = \mathbf{x}_{ij}^\top\boldsymbol{\beta} + \mathbf{w}_{ij}^\top\mathbf{b}_i, \\ \mathbf{b}_i &\sim N_q(\mathbf{0}, \mathbf{D}), \end{aligned} \tag{2}$$

where $\Phi(\cdot)$ denotes the cdf of the standard normal distribution, $N(0,1)$, $\mathbf{x}_{ij}^\top = (x_{ij1}, \ldots, x_{ijp})^\top$ and $\mathbf{w}_{ij}^\top = (w_{ij1}, \ldots, w_{ijq})^\top$ are covariates, $\boldsymbol{\beta}$ is the $p \times 1$ fixed effects, $\{\mathbf{b}_i\}_{i=1}^m$ are the $q \times 1$ random effects, $\mathbf{D}$ is a $q \times q$ unknown matrix relating to the correlation structure of $\mathbf{Y}_i$. This model can be alternatively written in terms of an underlying latent continuous variable $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{in_i})^\top$, such that

$$\begin{aligned} Y_{ij} = \mathbb{I}_{(0,\infty)}(Z_{ij}), \qquad \mathbf{Z}_i|\mathbf{b}_i &\sim N_{ni}(\boldsymbol{\mu}_i, \mathbf{I}_{n_i}), \\ \mathbf{b}_i &\sim N_q(\mathbf{0}, \mathbf{D}), \end{aligned}$$

where $\mathbf{I}_{ni}$ denotes the $n_i \times n_i$ identity matrix and

$$\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{in_i})^\top = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{W}_i\mathbf{b}_i,$$

with $\mathbf{X}_i = (\mathbf{x}_{i1} \ldots, \mathbf{x}_{in_i})^\top$, $\mathbf{W}_i = (\mathbf{w}_{i1} \ldots, \mathbf{w}_{in_i})^\top$ being $n_i \times p$ and $n_i \times q$ matrices, respectively. The observed-data likelihood for $\boldsymbol{\theta} = (\boldsymbol{\beta}, \mathbf{D})$ is

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{m} \int \phi_q(\mathbf{b}_i|\mathbf{0}, \mathbf{D})\psi_i^n(\boldsymbol{\beta}, \mathbf{b}_i)d\mathbf{b}_i,$$

where $\phi_q(.|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ stands for the pdf of the q–variate normal distribution with mean vector $\boldsymbol{\mu}$ and covariate matrix $\boldsymbol{\Sigma}$ and

$$\psi_i^n(\boldsymbol{\beta}, \mathbf{b}_i) = \prod_{j=1}^{n_i}[\Phi(\mu_{ij})]^{y_{ij}}[1 - \Phi(\mu_{ij})]^{1-y_{ij}}$$

and $\mathbf{y}_i = (y_{i1}, \ldots, y_{in_i})^\top$ denotes a realization of $\mathbf{Y}_i$.

Under normality, several methods have been proposed to efficiently compute ML estimates of the unknown variance parameters ($\mathbf{D}$), as well as ML estimates of the fixed effects ($\boldsymbol{\beta}$). For instance, a Monte Carlo EM (MCEM) algorithm with a Gibbs sampler at each E-step was proposed by McCulloch (1994). To fit more general models, McCulloch (1997) used a Metropolis - Hastings algorithm at each E-step in the MCEM. An approximated procedure based on the h-likelihood was proposed by Lee and Nelder (2006). Later, Tan et al. (2007) proposed a non-iterative importance sampling approach based on the inverse Bayes formula (IBF), where the first- and second-order moments of a truncated multivariate normal distribution were evaluated and associated with the MCEM algorithm. Another approach, proposed to use the SAEM algorithm (Delyon et al., 1999), was introduced by Meza et al. (2009). Alternatively, Bayesian inference can be carried out with Markov Chain Monte Carlo (MCMC) and implemented via Gibbs sampling (Albert and Chib, 1993). In the next section we present the proposed robust t-link for GLMM.

3.2 The t-GLMM for binary data

Pinheiro et al. (2001) proposed a robust hierarchical linear mixed-effects model in which the random effects and the within-subject errors have multivariate Student's t-distributions. In its formulation it is assumed that the mixture distributions for the two sources of variability in the model have the same shape and share the same parameters. We proceed as in Pinheiro et al. (2001) (see also Matos et al., 2013) by considering a generalization of the classic probit-GLMM for binary data defined in (2), as follows:

$$\begin{aligned} Y_{ij} = \mathbb{I}_{(0,\infty)}(Z_{ij}),\ \mathbf{Z}_i|\mathbf{b}_i, U_i = u_i &\sim N_{ni}(\boldsymbol{\mu}_i, u_i^{-1}\mathbf{I}_{n_i}), \\ \mathbf{b}_i|U_i = u_i &\sim N_q(\mathbf{0}, u_i^{-1}\mathbf{D}), \\ U_i &\sim \text{Gamma}(\nu/2, \nu/2). \end{aligned} \tag{3}$$

Using Lemma 1, given in the Appendix, the model defined in (3) is equivalent to the following representation

$$Pr(Y_{ij} = 1|\mathbf{b}_i) = T_1(\mu_{ij}|0, 1, \nu), \quad \mathbf{b}_i \sim t_q(\mathbf{0}, \mathbf{D}, \nu),$$

where $\mu_{ij} = \mathbf{x}_{ij}^\top \boldsymbol{\beta} + \mathbf{w}_{ij}^\top \mathbf{b}_i$.

The classic inference on the parameter vector $\boldsymbol{\theta}$ is based on the (observed-data) likelihood of $\boldsymbol{\theta}$ given the observed sample $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)$, which in this case is given by

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{m} \int t_q(\mathbf{b}_i|\mathbf{0}, \mathbf{D}, \nu)\psi_i^t(\boldsymbol{\beta}, \mathbf{b}_i) d\mathbf{b}_i, \tag{4}$$

where $t_q(.|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$ is as defined in Section 2 and

$$\psi_i^t(\boldsymbol{\beta}, \mathbf{b}_i) = \prod_{j=1}^{n_i} \left[T_1(\mu_{ij}|0, 1, \nu)\right]^{y_{ij}} \left[1 - T_1(\mu_{ij}|0, 1, \nu)\right]^{1-y_{ij}}. \tag{5}$$

Note that the likelihood (4) does not have a closed form expression because the model function is not linear in the random effect. To compute ML estimates of the unknown variance parameters, the EM algorithm is proposed. This algorithm relies on formulas for the mean and variance of a truncated multivariate t-distribution, which can be computed using available formulas.

On the other hand, although several authors have addressed parameter estimation of mixing variables in modeled regressions (see, for instance, Lange and Sinsheimer, 1993; Jamshidian, 1999) based on the EM algorithm and its variants. Fernandez and Steel (1999) discussed potential problems that may arise in the estimation of degrees of freedom, in particular for the Student's t-distribution. This is due to the apparent unboundedness of the likelihood function near the boundary of the parameter space, and hence the ML scheme as developed in Lange and Sinsheimer (1993) are questionable because they do not provide sufficient information on whether these estimates correspond to local, or global maxima. Interestingly, Lucas (1997) noted that only under fixed degrees of freedom do the parameter estimates behave robustly against extreme observations. A plausible (and simple) alternative is to assume that the parameter ν associated with the mixture variables U is known, which has been adopted in this work. Recent works in the context of elliptical distributions have considered the parameter ν to be known. See for instance, Meza et al. (2009) and Matos et al. (2013).

4 The EM algorithm

In this section, we first derive the M- and E- step for the proposed EM algorithm, and then we show how to use exact formulas instead of Monte Carlo (MC) sampling at each E-step. For the probit model, a MCEM algorithm was proposed by McCulloch (1994), with computational improvements given recently by Meza et al. (2009) and Tan et al. (2007). Following the notation of the former article, we treat both $\mathbf{b} = \{\mathbf{b}_i\}_{i=1}^m$, $\mathbf{Z} = \{\mathbf{Z}_i\}_{i=1}^m$ and $\mathbf{U} = \{U_i\}_{i=1}^m$ as missing data and $\mathbf{Y} = \{\mathbf{Y}_i\}_{i=1}^m$ as the observed data. Then, the joint density for the complete-data $\mathbf{Y}_{com} = \{\mathbf{Y}, \mathbf{Z}, \mathbf{b}, \mathbf{U}\}$ is

$$L(\boldsymbol{\theta}|\mathbf{Y}_{com}) = \prod_{i=1}^{m} \left[\phi_q(\mathbf{b}_i|\mathbf{0}, u_i^{-1}\mathbf{D})\phi_{n_i}(\mathbf{Z}_i|\boldsymbol{\mu}_i, u_i^{-1}\mathbf{I}_{n_i})G(u_i|\nu/2, \nu/2)\right], \tag{6}$$

The purpose of the M-step of the EM algorithm is to find the complete-data MLE of $\boldsymbol{\theta} = (\boldsymbol{\beta}, \mathbf{D})$ by maximizing the conditional expectation of the complete-data log-likelihood $\ell(\boldsymbol{\theta}|\mathbf{Y}_{com}) = \log L(\boldsymbol{\theta}|\mathbf{Y}_{com})$ given the observed data $\mathbf{Y}$ and the current estimate $\boldsymbol{\theta}^{(k)}$, given by

$$\begin{aligned} Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k)}) &= E[\ell(\boldsymbol{\theta}|\mathbf{Y}_{com})|\mathbf{Y}] = C - \frac{m}{2}log|\mathbf{D}| - \frac{1}{2}\sum_{i=1}^{m} tr\left(\mathbf{D}^{-1}E[U_i\mathbf{b}_i\mathbf{b}_i^\top|\mathbf{Y}_i]\right) \\ &\quad + \sum_{i=1}^{m}\left(\boldsymbol{\beta}^\top\mathbf{X}_i^\top\mathbf{X}_i\boldsymbol{\beta}E[U_i|\mathbf{Y}_i]\right) - 2\sum_{i=1}^{m}\left[\boldsymbol{\beta}^\top\mathbf{X}_i^\top\left(E[U_i\mathbf{Z}_i|\mathbf{Y}_i] - \mathbf{W}_iE[U_i\mathbf{b}_i|\mathbf{Y}_i]\right)\right] \end{aligned}$$

where C is a constant that is independent of the parameters. Thus, in the M-step we update $\boldsymbol{\beta}$, $\mathbf{D}$ through the following closed form expressions

$$
\begin{aligned}
\widehat{\boldsymbol{\beta}} &= (\sum_{i=1}^{m} E[U_i|\mathbf{Y}_i]\mathbf{X}_i^\top \mathbf{X}_i)^{-1} \sum_{i=1}^{m}(\mathbf{X}_i^\top E[U_i\mathbf{Z}_i|\mathbf{Y}_i] - \\
&\quad -\mathbf{W}_i E[U_i\mathbf{b}_i|\mathbf{Y}_i]) \quad \text{and} \qquad (7)\\
\widehat{\mathbf{D}} &= \frac{1}{m}\sum_{i=1}^{m} E[U_i\mathbf{b}_i\mathbf{b}_i^\top|\mathbf{Y}_i], \qquad (8)
\end{aligned}
$$

where $E[U_i|\mathbf{Y}_i]$, $E[U_i\mathbf{Z}_i|\mathbf{Y}_i]$, $E[U_i\mathbf{b}_i|\mathbf{Y}_i]$ and $E[U_i\mathbf{b}_i\mathbf{b}_i^\top|\mathbf{Y}_i]$ are expected values in $(U_i, \mathbf{b}_i, \mathbf{Z}_i)$ conditional on $\mathbf{Y}_i$, taken at the current parameter value $\boldsymbol{\theta}^{(k)} = (\boldsymbol{\beta}^{(k)}, \mathbf{D}^{(k)})$. It can be shown (see Appendix) that

$$
\begin{aligned}
E[U_i|\mathbf{Y}_i] &= \bar{\mathbf{Z}}_i^0, \ E[U_i\mathbf{Z}_i|\mathbf{Y}_i] = \bar{\mathbf{Z}}_i^1,\\
E[U_i\mathbf{b}_i|\mathbf{Y}_i] &= \boldsymbol{\Delta}_i(\bar{\mathbf{Z}}_i^1 - \bar{\mathbf{Z}}_i^0\mathbf{X}_i\boldsymbol{\beta}), \qquad (9)\\
E[U_i\mathbf{b}_i\mathbf{b}_i^\top|\mathbf{Y}_i] &= \boldsymbol{\Lambda}_i + \boldsymbol{\Delta}_i(\bar{\mathbf{Z}}_i^2 + \boldsymbol{\gamma}_i\boldsymbol{\gamma}_i^\top\bar{\mathbf{Z}}_i^0 - \bar{\mathbf{Z}}_i^1\boldsymbol{\gamma}_i^\top - \boldsymbol{\gamma}_i\bar{\mathbf{Z}}_i^{1^\top})\boldsymbol{\Delta}_i,\\
\mathbf{Z}_i|\mathbf{Y}_i &\sim t_{n_i}(\boldsymbol{\gamma}_i, \boldsymbol{\Omega}_i, \nu)\mathbb{I}_{\mathbb{B}_i}(\mathbf{Z}_i),
\end{aligned}
$$

where $\bar{\mathbf{Z}}_i^0 = E\left[\frac{\nu+n_i}{\nu+\delta_i}|\mathbf{Y}_i\right]$, $\bar{\mathbf{Z}}_i^1 = E\left[\frac{\nu+n_i}{\nu+\delta_i}\mathbf{Z}_i|\mathbf{Y}_i\right]$, $\bar{\mathbf{Z}}_i^2 = E\left[\frac{\nu+n_i}{\nu+\delta_i}\mathbf{Z}_i\mathbf{Z}_i^\top|\mathbf{Y}_i\right]$, $\delta_i = (\mathbf{Z}_i - \boldsymbol{\gamma}_i)^\top\boldsymbol{\Omega}_i^{-1}(\mathbf{Z}_i - \boldsymbol{\gamma}_i)$, $\boldsymbol{\Delta}_i = \mathbf{D}\mathbf{W}_i^\top\boldsymbol{\Omega}_i^{-1}$, $\boldsymbol{\Lambda}_i = \mathbf{D} - \mathbf{D}\mathbf{W}_i^\top\boldsymbol{\Omega}_i^{-1}\mathbf{W}_i\mathbf{D}$, $\boldsymbol{\Omega}_i = \mathbf{W}_i\mathbf{D}\mathbf{W}_i^\top + \mathbf{I}_{n_i}$, $\boldsymbol{\gamma}_i = \mathbf{X}_i\boldsymbol{\beta}$, and $\mathbb{B}_i = B_{i1}\times\ldots,\times B_{in_i}$, where B_{ij} is the interval $(0,\infty)$ if $y_{ij} = 1$ and the interval $(-\infty, 0]$ if $y_{ij} = 0$.

From (7)-(8), the E-step reduces to computation of $\bar{\mathbf{Z}}_i^0$, $\bar{\mathbf{Z}}_i^1$ and $\bar{\mathbf{Z}}_i^2$. From (9) it is clear that Proposition 1 cannot be used, since the components of the random vector $\mathbf{Z}_i|\mathbf{Y}_i$ are right or left truncated depending on $y_{ij}, j = 1, \ldots, n_i$. However, these quantities can be determined in closed form using a sequence of simple transformations, as follows.

(i) The first step is to standardize the components of $\mathbb{B}_i$, either as left- or right truncated. Let $\mathbf{A}_i$ be a diagonal matrix with diagonal elements equal to -1 or 1 depending on $B_{ij} = (0,\infty)$ or $B_{ij} = (-\infty, 0]$, respectively. Then, $\mathbf{U}_i \equiv \mathbf{A}_i\mathbf{Z}_i|\mathbf{Y}_i \sim Tt_{n_i}(\mathbf{A}_i\boldsymbol{\gamma}_i, \mathbf{A}_i\boldsymbol{\Omega}_i\mathbf{A}_i, \nu; \mathbb{C}_i)$, $\mathbb{C}_i = (-\infty, 0]^{n_i}$, that is, $\mathbf{U}_i$ follows a multivariate t-distribution $t_{n_i}(\mathbf{A}_i\boldsymbol{\gamma}_i, \mathbf{A}_i\boldsymbol{\Omega}_i\mathbf{A}_i, \nu)$ right truncated at $(-\infty, 0]^{n_i}$. This standardization facilitates the computation of $\bar{\mathbf{U}}_i^0 = E\left[\frac{\nu+n_i}{\nu+\delta_i^u}|\mathbf{Y}_i\right]$, $\bar{\mathbf{U}}_i^1 = E\left[\frac{\nu+n_i}{\nu+\delta_i^u}\mathbf{U}_i|\mathbf{Y}_i\right]$, $\bar{\mathbf{U}}_i^2 = E\left[\frac{\nu+n_i}{\nu+\delta_i^u}\mathbf{U}_i\mathbf{U}_i^\top|\mathbf{Y}_i\right]$, through the result given in Proposition 1 along with the computation of the first two moments of a truncated multivariate t-distribution with specific parameters, where $\delta_i^u = (\mathbf{U}_i - \mathbf{A}_i\boldsymbol{\gamma}_i)^\top(\mathbf{A}_i\boldsymbol{\Omega}_i\mathbf{A}_i)^{-1}(\mathbf{U}_i - \mathbf{A}_i\boldsymbol{\gamma}_i)$.

(ii) The second step is to note that $\bar{\mathbf{Z}}_i^0 = \bar{\mathbf{U}}_i^0$, $\bar{\mathbf{Z}}_i^1 = \mathbf{A}_i^{-1}\bar{\mathbf{U}}_i^1$ and $\bar{\mathbf{Z}}_i^2 = \mathbf{A}_i^{-1}\bar{\mathbf{U}}_i^2\mathbf{A}_i^{-1}$, since $\delta_i^u = \delta_i = (\mathbf{Z}_i - \boldsymbol{\gamma}_i)^\top\boldsymbol{\Omega}_i^{-1}(\mathbf{Z}_i - \boldsymbol{\gamma}_i)$.

When ν goes to ∞, we have an interesting EM-type algorithm for the probit model defined in (2).

4.1 Estimation of the likelihood

The observed-data likelihood for $\boldsymbol{\theta} = (\boldsymbol{\beta}, \mathbf{D})$ is given by (4)-(5). This integral can be conveniently computed via an importance sampling scheme for any continuous distribution $\widetilde{\pi}$ (with a support larger than that of $\pi(\mathbf{b}_i, \boldsymbol{\theta}) \equiv t_q(\mathbf{b}_i|\mathbf{0}, \mathbf{D}, \nu)$). Thus, Eq. (4) can be represented as

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{m} \int \psi_i^t(\boldsymbol{\beta}, \mathbf{b}_i)\frac{\pi(\mathbf{b}_i, \boldsymbol{\theta})}{\widetilde{\pi}(\mathbf{b}_i, \boldsymbol{\theta})}\widetilde{\pi}(\mathbf{b}_i, \boldsymbol{\theta})d\mathbf{b}_i,$$

where $\psi_i^t(\boldsymbol{\beta}, \mathbf{b}_i)$ as in (5). Therefore,$\ell(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta})$ can be estimated without additional computational by

$$\widehat{\ell(\boldsymbol{\theta})} \doteq \sum_{i=1}^{m} \log\left[\frac{1}{K}\sum_{l=1}^{K}[\psi_i(\boldsymbol{\beta}, \mathbf{b}_i^{(l)})\frac{\pi(\mathbf{b}_i^{(l)}, \boldsymbol{\theta})}{\widetilde{\pi}(\mathbf{b}_i^{(l)}, \boldsymbol{\theta})}]\right].$$

where $\mathbf{b}_1, \ldots, \mathbf{b}_l, \ldots, \mathbf{b}_K$ are draws from $\widetilde{\pi}(\mathbf{b}_i, \boldsymbol{\theta})$. An efficient choice for $\widetilde{\pi}$ consists of the conditional distribution of $\mathbf{b}_i$ given the data $\mathbf{y}_i$, $i = 1, \ldots, m$ (Robert et al., 1999).

Another approach to calculate the likelihood is to consider the *adaptIntegrate* function of the *cubature* package (Johnson and Narasimhan, 2011) available in R. This function is appropriate for numerical multivariate integration and will be used to calculate the log-likelihood in the application section.

The log-likelihood can be used to monitor the convergence of the EM-algorithm. In practice, the iterations are repeated until some distance involving two successive evaluations of the actual log-likelihood $\ell(\boldsymbol{\theta})$, like $||\ell(\widehat{\boldsymbol{\theta}}^{(k+1)}) - \ell(\widehat{\boldsymbol{\theta}}^{(k)})||$ or $||\ell(\widehat{\boldsymbol{\theta}}^{(k+1)})/\ell(\widehat{\boldsymbol{\theta}}^{(k)}) - 1||$, $k = 0, 2, ...$, is small enough. In addition, model selections based on the observed likelihood can be done using appropriate likelihood ratio tests (AIC and BIC).

4.2 Standard error approximation

Denote the ML estimates from the EM-algorithm by $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\beta}}, \widehat{\mathbf{D}})$. Under some general regularity conditions, we follow McLachlan and Krishnan (1997) to provide an information-based method to obtain the asymptotic covariance of the ML estimates for t-GLMM for correlated binary data. We define $I_o(\widehat{\boldsymbol{\theta}}|\mathbf{Y}) = \sum_{i=1}^{m} \widehat{\mathbf{u}}_i \widehat{\mathbf{u}}_i^{\top}$ to be the observed information matrix, where $\widehat{\mathbf{u}}_i = E\left[\frac{\partial \ell_i(\boldsymbol{\theta}|\mathbf{Y}_{comp})}{\partial \boldsymbol{\theta}}|\mathbf{y}_i\right]|_{\widehat{\boldsymbol{\theta}}} = \frac{\partial Q_i(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}}|_{\widehat{\boldsymbol{\theta}}}$, with $\ell_i(\boldsymbol{\theta}|\mathbf{Y}_{comp})$ being the complete - data log-likelihood formed from the single observation $\mathbf{y}_i$, $i = 1, \ldots, m$.

Explicit expressions for the elements of $\widehat{\mathbf{u}}_i$ are

$$\begin{aligned}
\frac{\partial Q_i(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})}{\partial \boldsymbol{\beta}}|_{\widehat{\boldsymbol{\theta}}} &= \mathbf{X}_i^{\top}(\widehat{U_i\mathbf{Z}_i} - \mathbf{W}_i\widehat{U_i\mathbf{b}_i}) - (\mathbf{X}_i\mathbf{X}_i^{\top})\widehat{U}_i\widehat{\boldsymbol{\beta}} \\
\frac{\partial Q_i(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}})}{\partial d_r}|_{\widehat{\boldsymbol{\theta}}} &= -\frac{1}{2}\left[tr\left(\widehat{\mathbf{D}}^{-1}\frac{\partial \widehat{\mathbf{D}}}{\partial d_r}\widehat{\mathbf{D}}^{-1}(\widehat{\mathbf{D}} - \widehat{\mathbf{b}_i\mathbf{b}_i U_i})\right)\right]
\end{aligned}$$

where $\widehat{U}_i = E[U_i|\mathbf{Y}_i]\big|_{\widehat{\boldsymbol{\theta}}} = \bar{\mathbf{Z}}_i^0\big|_{\widehat{\boldsymbol{\theta}}}$, $\widehat{U_i\mathbf{Z}_i} = E[U_i\mathbf{Z}_i|\mathbf{Y}_i]\big|_{\widehat{\boldsymbol{\theta}}} = \bar{\mathbf{Z}}_i^1\big|_{\widehat{\boldsymbol{\theta}}}$, $\widehat{U_i\mathbf{b}_i} = E[U_i\mathbf{b}_i|\mathbf{Y}_i]\big|_{\widehat{\boldsymbol{\theta}}}$, $\widehat{\mathbf{b}_i\mathbf{b}_i U_i} = E[\mathbf{b}_i\mathbf{b}_i^{\top}U_i|\mathbf{Y}_i]\big|_{\widehat{\boldsymbol{\theta}}}$, d_r are the distinct elements of the square matrix $\mathbf{D}$. Standard errors are equal to the square roots of the diagonal elements of the inverse of the estimated information matrix $I_o(\widehat{\boldsymbol{\theta}}|\mathbf{Y})$.

4.3 Estimation of the random effects

In this section, we consider an empirical Bayes inference for the random effects, which is useful for examining subject-specific quantities of interest. From the conditional distribution of $\mathbf{b}_i$ given $(\mathbf{Y}_i, \mathbf{Z}_i, U_i)$ (see the Appendix), we have:

$$\begin{aligned}
f(\mathbf{b}_i|\mathbf{Y}_i, \mathbf{Z}_i, u_i, \boldsymbol{\theta}) &= f(\mathbf{b}_i|\mathbf{Z}_i, u_i, \boldsymbol{\theta}) \\
&= \phi_q(\mathbf{b}_i|\boldsymbol{\Delta}_i(\mathbf{Z}_i - \mathbf{X}_i^{\top}\boldsymbol{\beta}), u_i^{-1}\boldsymbol{\Lambda}_i),
\end{aligned}$$

where, $\boldsymbol{\Delta}_i = \mathbf{D}\mathbf{W}_i^{\top}\boldsymbol{\Omega}_i^{-1}$, $\boldsymbol{\Lambda}_i = \mathbf{D} - \mathbf{D}\mathbf{W}_i^{\top}\boldsymbol{\Omega}_i^{-1}\mathbf{W}_i\mathbf{D}$ and $\boldsymbol{\Omega}_i = \mathbf{W}_i\mathbf{D}\mathbf{W}_i^{\top} + \mathbf{I}_{n_i}$, $i = 1, \ldots, m$.

Thus, the minimum mean-squared error (MSE) predictor of $\mathbf{b}_i$, obtained by the conditional mean of $\mathbf{b}_i$ given $\mathbf{Y}_i = \mathbf{y}_i$, is

$$\begin{aligned}
\hat{\mathbf{b}}_i = E\left[\mathbf{b}_i|\mathbf{Y}_i, \boldsymbol{\theta}\right] &= E\left[E(\mathbf{b}_i|\mathbf{Y}_i, \mathbf{Z}_i, \boldsymbol{\theta})|\mathbf{Y}_i, \boldsymbol{\theta}\right] \\
&= \boldsymbol{\Delta}_i[E(\mathbf{Z}_i|\mathbf{Y}_i, \boldsymbol{\theta}) - \mathbf{X}_i^{\top}\boldsymbol{\beta}],
\end{aligned}$$

where $E[\mathbf{Z}_i|\mathbf{Y}_i, \boldsymbol{\theta}]$ is the expected value of the multivariate truncated t-distribution given in (9).

5 Applications

5.1 Simulation study

Here we present a simulation to study the performance of our proposed method. The main goal of this simulation study is to investigate the effects on the parameters inference when the common normality assumption is violated for the link, as well as, for the random effects. Moreover, the simulation has the purpose of investigating if the metric used for model comparison (AIC and BIC) determines the correct model for the simulated data.

The simulation study mimics the structure of the real data analysis presented in Subsection 5.2. We define a balanced design for the response, with $j = 1, \ldots, 6$, to represent the number of repeated measurements. We consider the following GLMM for binary data:

$$
\begin{aligned}
Y_{ij} = \mathbb{I}_{(Z_{ij}>0)},\ \mathbf{Z}_i|\mathbf{b}_i, U_i = u_i &\sim N_{ni}(\boldsymbol{\mu}_i, u_i^{-1}\mathbf{I}_{n_i}),\\
\mathbf{b}_i|U_i = u_i &\sim N_q(\mathbf{0}, u_i^{-1}\mathbf{D}),\\
U_i &\sim F(\nu),
\end{aligned}
$$

where $\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{i6})^\top$, with $\mu_{ij} = \beta_0 + \beta_1 X_{i1} + b_{1i} + b_{2i}X_{i2}$ for all j. A dichotomous covariate X_1 mimicking *gender* is generated for the fixed effects and a standard normal random covariate X_2 is created to mimic a scaled continuous variables, e.g. *age*, in the random effects; $F(\nu)$ is a positive distribution. Finally, we set $\beta_0 = 1$, $\beta_1 = 0.8$ and define the matrix $\mathbf{D} = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$ to represent the correlation structure between the observation associates with the i-th individual.

To fulfill our objectives, we present three simulation scenarios with different choices for $F(\nu)$:

1. Probit: The data have a probit link and normal random effects. So, $U = 1$ is a degenerate distribution with $P(U = 1) = 1$.

2. T-link: The data have a Student's t-link and Student's t random effects with degrees of freedom $\nu = 4$. So, $U \sim \text{Gamma}(\nu/2, \nu/2)$.

3. Contaminated-link: The data have a contaminated normal link and contaminated normal random effects with degrees of freedom $\nu = (\nu_1, \nu_2)^\top =$

$(0.1, 0.1)^{\top}$. So, $U = \begin{cases} \nu_2 & \text{with prob} \quad \nu_1 \\ 1 & \text{with prob} \quad 1-\nu_1 \end{cases}$.

For all scenarios, sample sizes of 100, 250 and 500 were generated. Under the three scenarios, we fitted the probit-GLMM (Subsection 3.1) and the t-GLMM with 4 degrees of freedom (Subsection 3.2) models to analyze the characteristics of each fitting model when the true model is one of them and when none is the true generating model. This way, we have 9 different simulation settings with 1000 simulated datasets under each setting. For each simulation, the parameter estimates as well as the AIC and BIC were recorded.

Table 1 presents the summary statistics for β_0 and β_1 (the fixed-effects parameters) assuming that the true model is the probit-GLMM for the 3 proposed sample sizes. In this table, MC Mean denotes the arithmetic average of the 1000 estimates given by $\sum_{j=1}^{1000} \hat{\gamma}_j/1000$ and MC Sd is the arithmetic average of the 1000 standard deviations of the estimates of the parameters given by $\sum_{j=1}^{1000} sd(\hat{\gamma}_j)/1000$, where $\gamma = \beta_0$ or β_1. In addition, we also estimate the MC coverage of β_0 and β_1, i.e., the proportion of times the 95% confidence interval includes the true value of the fixed effects.

From Table 1, we can see that the probit model outperforms the t-link model parameter estimation at all levels, as expected. It is clear that the probit-GLMM and the t-GLMM are able to recover the true parameter values and the estimation improves as the sample size increases. Moreover, Table 1 also provides the average values of the approximate standard errors of the EM estimates obtained through the information-based method, described in Subsection 4.2 (IM Sd), and the Monte Carlo standard deviation (MC Sd) for the parameters. As can be seen, the theoretically estimated standard errors are relatively close to the Monte Carlo standard deviation estimates obtained empirically. This result shows that the proposed asymptotic approximation for the variances of the fixed effects is reliable for both probit-GLMM and t-GLMM models. We also present the arithmetic averages (MC AIC and MC BIC) of the model comparison criteria mentioned earlier. As can be seen, the MC AIC and MC BIC correctly select the probit-GLMM as the preferred model.

We continue to investigate the performance of our proposed methods under scenario 2, in which the t-GLMM is the true model. Table 2 shows that the t-GLMM fits the data better, as expected, since the true generating model was a t-GLMM. However, when the data are not generated from a probit, the probit-GLMM does

Table 1: Results based on 1000 simulated probit samples. MC mean, MC Sd (in parentheses) and MC Coverage are the respective mean estimates, standard deviations and coverage proportion average from fitting the probit-GLMM and the t-GLMM. IM SE are the average values of the approximate standard errors obtained through the information-based method. MC AIC and MC BIC are the arithmetic averages of the respective model comparison measures.

			Simulated probit data			
Sample Size	Fit		β_0	β_1	MC AIC	MC BIC
100	probit	MC Mean	1.006	0.843	554.59	570.22
		IM SE	0.202	0.478		
		MC Sd	(0.171)	(0.399)		
		MC Coverage	98%	98%		
	t-link	MC Mean	1.062	0.933	556.23	571.89
		IM SE	0.217	0.612		
		MC Sd	(0.179)	(0.428)		
		MC Coverage	99%	99%		
250	probit	MC Mean	1.007	0.815	1371.69	1392.81
		IM SE	0.127	0.280		
		MC Sd	(0.125)	(0.246)		
		MC Coverage	97%	97%		
	t-link	MC Mean	1.061	0.899	1375.99	1397.12
		IM SE	0.136	0.363		
		MC Sd	(0.133)	(0.272)		
		MC Coverage	96%	99%		
500	probit	MC Mean	1.005	0.820	2734.52	2742.98
		IM Sd	0.090	0.195		
		MC SE	(0.106)	(0.172)		
		MC Coverage	91%	97%		
	t-link	MC Mean	1.060	0.902	2759.80	2768.27
		IM SE	0.096	0.253		
		MC Sd	(0.114)	(0.196)		
		MC Coverage	88%	98%		

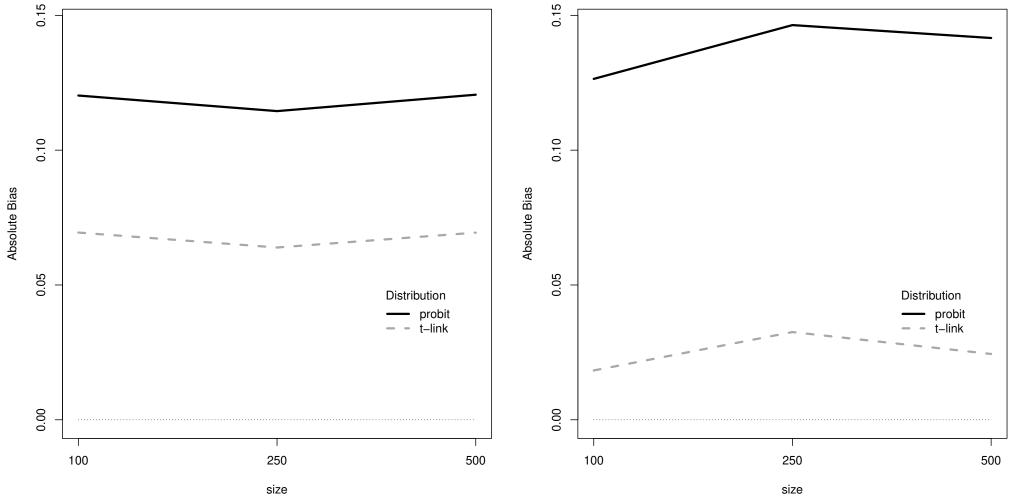

Figure 1: Simulation study. Comparison between the probit-GLMM and t-GLMM under the scenario 3. (Left panel) represents the bias of β_0. (Right panel) represents the bias of β_1.

not adjust to the t-GLMM as well as the t-link model did in the previous scenario. This is a clear indication of the robustness of the t-GLMM. Equivalent to the results obtained in Table 1, we can conclude that the standard errors are adequately calculated. Like to what occurred with the probit scenario, the MC AIC and MC BIC still select the correct generating model.

Now we introduce scenario 3, where the main objective is to study how the probit-GLMM and the t-GLMM behave when neither of them are the true generating model. Table 3 indicates that the t-GLMM adjusts better to the contaminated normal data. This finding is more evident in Figure 1. Another verification of the robustness of the t-GLMM is that both MC AIC and MC BIC select it as the preferred model. Therefore, the t-link is more robust to deviations from the model assumptions and fits better than the probit-GLMM when neither is the true generating model. As seen in the previous scenarios the IM SE estimates are close to the MC Sd.

Finally, one of our main goals of this simulation study is to show is the ability of the model to recover the variance-covariance structure of the random effects $\mathbf{b}$, that is, the components of the matrix $\mathbf{D}$. Table 4 presents the analysis of the elements of the matrix $\mathbf{D}$ of all 9 scenarios. From Table 4, it is possible to see that both the probit and t-link models appropriately estimate the true values, improving their

Table 2: Results based on 1000 simulated t-link samples. MC mean, MC Sd (in parentheses) and MC Coverage are the respective mean estimates, standard deviations and coverage proportion average from fitting probit-GLMM and the t-GLMM models. IM SE are the average values of the approximate standard errors obtained through the information-based method. MC AIC and MC BIC are the arithmetic averages of the respective model comparison measures.

			Simulated t-link data			
Sample Size	Fit		β_0	β_1	MC AIC	MC BIC
100	probit	MC Mean	0.952	0.740	576.46	592.10
		IM SE	0.207	0.467		
		MC Sd	(0.183)	(0.423)		
		MC Coverage	96%	97%		
	t-link	MC Mean	1.004	0.831	574.23	589.86
		IM SE	0.221	0.580		
		MC Sd	(0.192)	(0.452)		
		MC Coverage	98%	98%		
250	probit	MC Mean	0.955	0.723	1426.14	1447.27
		IM SE	0.131	0.276		
		MC Sd	(0.119)	(0.239)		
		MC Coverage	94%	96%		
	t-link	MC Mean	1.004	0.813	1420.96	1442.09
		IM SE	0.140	0.345		
		MC Sd	(0.126)	(0.261)		
		MC Coverage	97%	99%		
500	probit	MC Mean	0.956	0.717	2846.13	2871.42
		IM SE	0.092	0.191		
		MC Sd	(0.106)	(0.177)		
		MC Coverage	88%	95%		
	t-link	MC Mean	1.004	0.809	2836.01	2861.30
		IM SE	0.098	0.240		
		MC Sd	(0.117)	(0.200)		
		MC Coverage	91%	99%		

Table 3: Results based on 1000 simulated probit-contaminated normal samples. MC mean, MC Sd (in parentheses) and MC Coverage are the respective mean estimates, standard deviations and coverage proportion average from fitting probit and the t-link models. IM SE are the average values of the approximate standard errors obtained through the information-based method. MC AIC and MC BIC are the arithmetic averages of the respective model comparison measures.

			Simulated contaminated-link data			
Sample Size	Fit		β_0	β_1	MC AIC	MC BIC
100	probit	MC Mean	0.880	0.699	592.13	607.77
		IM SE	0.192	0.430		
		MC Sd	(0.164)	(0.365)		
		MC Coverage	92%	96%		
	t-link	MC Mean	0.931	0.785	589.86	605.49
		IM SE	0.206	0.535		
		MC Sd	(0.172)	(0.393)		
		MC Coverage	96%	98%		
250	probit	MC Mean	0.886	0.683	1463.82	1484.95
		IM SE	0.121	0.257		
		MC Sd	(0.116)	(0.241)		
		MC Coverage	84%	93%		
	t-link	MC Mean	0.936	0.774	1458.92	1480.05
		IM SE	0.130	0.323		
		MC Sd	(0.124)	(0.264)		
		MC Coverage	93%	97%		
500	probit	MC Mean	0.879	0.687	2921.92	2947.20
		IM SE	0.085	0.179		
		MC Sd	(0.095)	(0.171)		
		MC Coverage	68%	90%		
	t-link	MC Mean	0.931	0.780	2911.85	2937.13
		IM SE	0.092	0.226		
		MC Sd	(0.105)	(0.197)		
		MC Coverage	84%	97%		

performance as the sample size increases. Moreover, as observed for the estimates of the fixed effects, the theoretically estimated standard errors for the dependence components are close to the Monte Carlo standard deviation estimates. This observation supports the ability of the method to properly estimate the standard errors for all parameters.

5.2 Analysis of real data

In this section, we analyze the dataset presented by Hjsgaard et al. (2005). The data come from a study of respiratory illness in 111 patients in a balanced design with 4 visits for each patient from two clinical centers. For each of the examination visits, the respiratory state of a patient was classified as good (= 1) or poor (= 0). The patients were randomized to receive either active treatment or placebo. Here, we assume that the probability $Pr(Y_{ij} = 1|b_i)$ of respiratory infection of the i-th patient at the j-th visit, $i = 1, \ldots, 111$, $j = 1, \ldots, 4$, is assumed to be equal to:

$$F(\beta_0 \quad + \quad \text{treat}\, \beta_1 + \text{gender}\, \beta_2 + \text{center}\, \beta_3 + \text{baseline}\, \beta_4 + age\, b_i), \tag{10}$$

where "F" is the probit or t-link; "treat" is the treatment of patient, 1 for active treatment and 0 for placebo; "gender = 0" for female gender and of 1 for male; "center = 0" for the first center and of 1 for the second; and "baseline" is the response at the first visit. The values of the covariates were constant for the visits. The element of the matrix $\mathbf{W}$ is $w_{ij} = \text{age}$, that is, the age of the patients included in the study centered around 31 years old (the median value). The random effects imply that the age effects are patient-specific (d_{11}).

In Figure 2, we plot the log-likelihood of the t-link model for different degrees of freedom. We notice that the smaller the degrees of freedom, the better is the fit. Using this result, we fixed $\nu = 4$ for our analysis.

Table 4: Monte Carlos estimates of the variance components based on 1000 simulated samples. The true elements are $d_{11} = 2$, $d_{12} = 1$ and $d_{22} = 1$. MC mean, MC Sd (in parentheses) and MC Coverage are the respective mean estimates, standard deviations and coverage proportion average from fitting probit and the t-link models. IM SE are the average values of the approximate standard errors obtained through the information-based method.

			probit			t-link		
Scenario	**D**		100	250	500	100	250	500
	$\hat{d}_{11}$	MC Mean	2.095	2.054	2.060	1.942	1.896	1.903
		IM SE	0.664	0.476	0.335	0.750	0.453	0.320
		MC Sd	(0.786)	(0.386)	(0.303)	(0.641)	(0.368)	(0.285)
		MC Coverage	97%	98%	98%	94%	95%	94%
	$\hat{d}_{12}$	MC Mean	1.025	0.997	0.993	1.051	1.023	1.021
probit		IM SE	0.556	0.334	0.234	0.538	0.326	0.229
		MC Sd	(0.448)	(0.262)	(0.203)	(0.443)	(0.258)	(0.200)
		MC Coverage	98%	98%	97%	97%	98%	97%
	$\hat{d}_{22}$	MC Mean	1.074	1.065	1.046	1.060	1.052	1.035
		IM SE	0.486	0.298	0.206	0.487	0.299	0.207
		MC Sd	(0.412)	(0.254)	(0.192)	(0.406)	(0.251)	(0.192)
		MC Coverage	97%	98%	97%	96%	97%	97%
	$\hat{d}_{11}$	MC Mean	2.245	2.244	2.228	2.094	2.092	2.079
		IM SE	0.830	0.514	0.358	0.793	0.491	0.343
		MC Sd	(0.685)	(0.449)	(0.320)	(0.664)	(0.432)	(0.308)
		MC Coverage	98%	99%	97%	97%	97%	98%
	$\hat{d}_{12}$	MC Mean	0.989	0.978	0.970	1.016	1.007	1.000
t-link		IM SE	0.583	0.348	0.241	0.567	0.339	0.236
		MC Sd	(0.448)	(0.294)	(0.212)	(0.443)	(0.289)	(0.210)
		MC Coverage	98%	97%	96%	98%	98%	97%
	$\hat{d}_{22}$	MC Mean	1.106	1.076	1.061	1.090	1.064	1.053
		IM SE	0.493	0.300	0.208	0.491	0.300	0.208
		MC Sd	(0.428)	(0.279)	(0.200)	(0.422)	(0.276)	(0.202)
		MC Coverage	97%	98%	98%	97%	97%	97%
	$\hat{d}_{11}$	MC Mean	1.922	1.902	1.908	1.804	1.785	1.793
		IM SE	0.693	0.426	0.300	0.667	0.410	0.289
		MC Sd	(0.599)	(0.365)	(0.264)	(0.583)	(0.357)	(0.256)
		MC Coverage	98%	99%	97%	97%	97%	98%
	$\hat{d}_{12}$	MC Mean	0.882	0.855	0.865	0.907	0.883	0.893
contaminated		IM SE	0.497	0.298	0.209	0.479	0.291	0.206
normal		MC Sd	(0.403)	(0.251)	(0.172)	(0.398)	(0.251)	(0.171)
		MC Coverage	98%	97%	96%	98%	98%	97%
	$\hat{d}_{22}$	MC Mean	0.995	0.963	0.965	0.988	0.959	0.963
		IM SE	0.438	0.265	0.186	0.439	0.266	0.188
		MC Sd	(0.389)	(0.242)	(0.166)	(0.386)	(0.242)	(0.167)
		MC Coverage	97%	98%	98%	97%	97%	97%

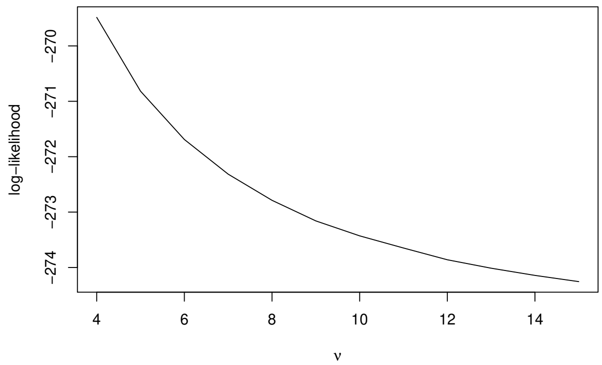

Figure 2: Plot of the profile log-likelihood ranging the degrees of freedom, ν.

Table 5: ML estimates for fitting the respiratory infection data with probit and t-link models.

		probit		t-link	
Variable	Parameter	MLE	SE	MLE	SE
intercept	β_0	-1.0992	0.5151	-1.1126	0.5965
treat	β_1	-0.9153	0.2524	-1.0589	0.2944
gender	β_2	0.1278	0.2970	0.0873	0.3447
center	β_3	0.5700	0.2415	0.6175	0.2751
baseline	β_4	1.6202	0.2650	1.7613	0.3061
covariance	d_{11}	0.0163	0.0061	0.0134	0.0053
AIC		575.5527		562.9703	
BIC		606.9342		594.3517	

Table 5 shows the ML estimates where the probit and t-link models were considered. At the E-step of the EM algorithm, we fixed $\boldsymbol{\beta}^{(0)} = (0, 0, 0, 0, 0)^\top$ and $\mathbf{D}^{(0)} = 1$ as the initial values. The maximum likelihood estimates for both models are relatively close. From the t-link column we can see that respiratory illness is related to treatment, center and baseline. Based on AIC and BIC, presented on Table 5, the t-link model is clearly the preferred one.

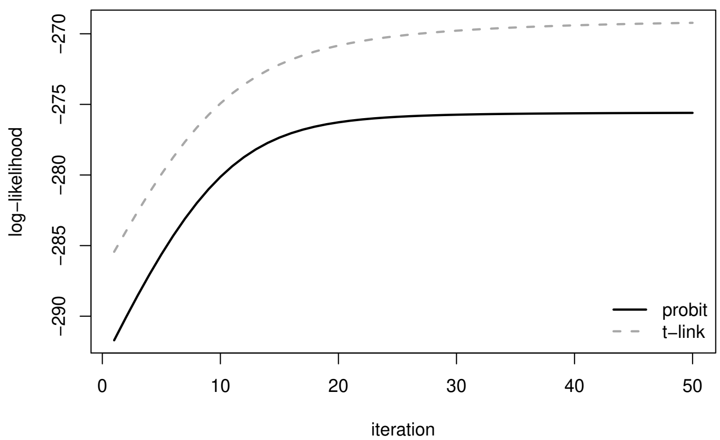

Figure 3: The comparison of convergence between two EM-type algorithms by plotting the log-likelihood values against the EM iteration.

To assess the convergence of the proposed EM, we compare the performance of the probit and t-link log-likelihood values with respect to the iterations of the algorithm. The results are illustrated in Figure 3. It shows that both models succeed in the convergence of the EM-type algorithm and that the t-link model clearly presents the best result for this dataset.

To obtain a sense of variation by the age random effect, we compute the empirical Bayes estimates of b_i, as defined in Section 4.3. The results are shown in Figure 4. Among those who received the active treatment, the greatest effect of age occurs for female patients in both models. For those who received placebos, the difference in gender is not substantial. It is worth noting, nonetheless, that in order to make strong conclusions about these effects across sex, new studies need to be carried out because of the highly unbalanced proportion of males and females in the study.

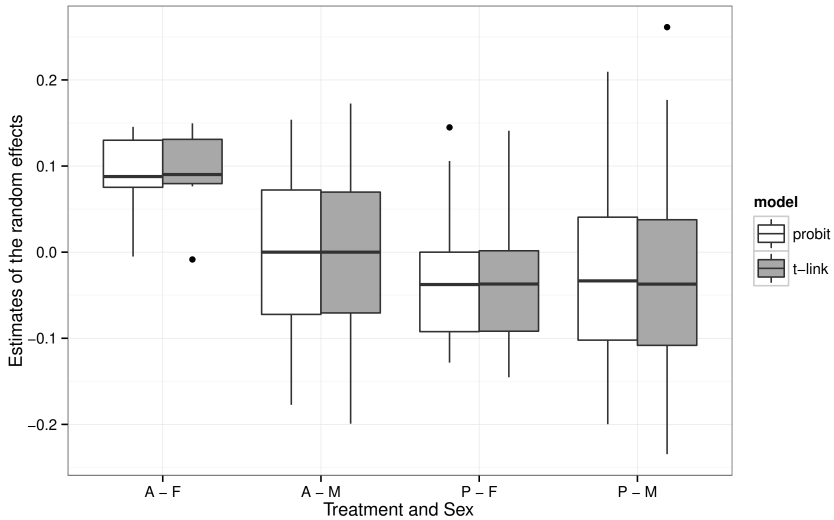

Figure 4: Empirical Bayes estimates of the random effects from both models across type of treatment and sex. The labels are: A=active treatment; P=placebo treatment; F=female; M=male.

6 Conclusions

In this article we developed an exact EM algorithm for correlated binary data with t-link. The algorithm has a closed-form expression for the E- step, based on formulas for the mean and variance of the truncated multivariate t-distribution. The computation uses existing functions for the multivariate t cumulative distribution function. Our applications showed that the t-link outperforms the typical probit-link in GLMM for binary data. The likelihood function is derived at no additional computational cost, paving the way for model selection procedures. As an additional benefit, the EM likelihood sequence is monotonic and the difficulties in assessing convergence which face MCMC algorithms are avoided. We believe this paper introduces a novel method and should also yield satisfactory results in other areas where truncated multivariate distributions appear frequently, for instance, Tobit models, item response theory models, among others. These extensions are currently under investigation. Finally, the proposed EM algorithm has been coded and implemented in an R script and is available from us on request.

Appendix

Proof of Proposition 1: First note that if $\mathbf{X} \sim t_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$, then we can write

$$\left(\frac{\nu+p}{\nu+\delta}\right)^r t_p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu) = c_p(\nu, r) t_p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}^*, \nu+2r).$$

It follows that

$$\begin{aligned} E\left[\left(\frac{\nu+p}{\nu+\delta}\right)^r \mathbf{X}^{(k)}\right] &= c_p(\nu, r)\frac{T_p(\mathbf{a}|\boldsymbol{\mu}, \boldsymbol{\Sigma}^*, \nu+2r)}{T_p(\mathbf{a}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)} \\ &\quad \times E\left[\mathbf{X}^{(k)}|\mathbf{X} \le \mathbf{a}\right], \end{aligned}$$

which concludes the proof.

Lemma 6.1. *If $U \sim Gamma(\alpha, \beta)$, then for any vector $\mathbf{B} \in \mathbb{R}^p$ and a $p \times p$ positive definite matrix $\boldsymbol{\Sigma}$,*

$$E[\Phi_p(\mathbf{B}\sqrt{U}|\mathbf{0}, \boldsymbol{\Sigma})] = T_p(\sqrt{\frac{\alpha}{\beta}}\mathbf{B}|\mathbf{0}, \boldsymbol{\Sigma}, 2\alpha),$$

Proof. If $\mathbf{V} \sim N_p(\mathbf{0}, \boldsymbol{\Sigma})$; then

$$\begin{aligned} E[\Phi_p(\mathbf{B}\sqrt{U}|\mathbf{0}, \boldsymbol{\Sigma})] &= E_U\left[P(\mathbf{V} \le \mathbf{B}\sqrt{u}|U=u)\right] \\ &= E_U[P(\frac{\mathbf{V}}{(u\beta/\alpha)^{1/2}} \le \sqrt{\frac{\alpha}{\beta}}\mathbf{B}|U=u)] \\ &= P(\mathbf{T} \le \sqrt{\frac{\alpha}{\beta}}\mathbf{B}), \end{aligned}$$

where, clearly $\mathbf{T} = \dfrac{\mathbf{V}}{(U\beta/\alpha)^{1/2}}$ has a multivariate Student's t-distribution, which concludes the proof. $\square$

Details of the EM Algorithm:

Treat $\mathbf{b} = \{\mathbf{b}_i\}_{i=1}^m$, $\mathbf{Z} = \{\mathbf{Z}_i\}_{i=1}^m$ and $\mathbf{U} = \{U_i\}_{i=1}^m$ as missing data. From the definition of the latent variable $\mathbf{Z}$, we have $\{\mathbf{Y}, \mathbf{Z}\} = \mathbf{Z}$. Then, the joint density for the complete-data $\mathbf{Y}_{com} = \{\mathbf{Y}, \mathbf{Z}, \mathbf{b}, \mathbf{U}\}$ is

$$\begin{aligned} f(\mathbf{Y}_{com}|\boldsymbol{\theta}) &= \prod_{i=1}^m f(\mathbf{b}_i|u_i, \mathbf{D}) f(\mathbf{Z}_i|\mathbf{b}_i, u_i, \beta) h(u_i|\nu) \\ &= \prod_{i=1}^m \phi_q(\mathbf{b}_i|\mathbf{0}, u_i^{-1}\mathbf{D})\phi_{n_i}(\mathbf{Z}_i|\boldsymbol{\mu}_i, u_i^{-1}\mathbf{I}_{n_i}) \times \\ &\times \; G(u_i|\nu/2, \nu/2). \end{aligned} \tag{11}$$

To complete the demonstration about how to employ the EM-type algorithm for ML estimation of the t-GLMM, it is necessary to derive the four conditional expectations of the complete-data sufficient statistics: $E[U_i|\mathbf{Y}_i]$, $E[U_i\mathbf{Z}_i|\mathbf{Y}_i]$, $E[U_i\mathbf{b}_i|\mathbf{Y}_i]$ and
$E[U_i\mathbf{b}_i\mathbf{b}_i^\top|\mathbf{Y}_i]$. To calculate them, we first derive the conditional predictive distribution of the missing data, which is given by:

$$\begin{aligned} f(\mathbf{b},\mathbf{Z},\mathbf{U}|\mathbf{Y},\boldsymbol{\theta}) &= f(\mathbf{Z}|\mathbf{Y},\mathbf{b},\mathbf{u},\boldsymbol{\theta})f(\mathbf{b}|\mathbf{Y},\mathbf{u},\boldsymbol{\theta})\times \\ &\times\ f(\mathbf{u}|\mathbf{Y},\boldsymbol{\theta}) = f(\mathbf{b}|\mathbf{Y},\mathbf{Z},\mathbf{u},\mathbf{D})\times \\ &\times\ f(\mathbf{u}|\mathbf{Y},\mathbf{Z},\boldsymbol{\theta})f(\mathbf{Z}|\mathbf{Y},\boldsymbol{\theta}). \end{aligned} \tag{12}$$

Since $f(\mathbf{b}|\mathbf{Y},\mathbf{Z},\mathbf{u},\boldsymbol{\theta})$ is proportional to (11), we obtain the following result:

$$\begin{aligned} f(\mathbf{b}|\mathbf{Y},\mathbf{Z},\mathbf{u},\boldsymbol{\theta}) &= \prod_{i=1}^m f(\mathbf{b}_i|\mathbf{Y}_i,\mathbf{Z}_i,u_i,\boldsymbol{\theta}) \\ &= \prod_{i=1}^m f(\mathbf{b}_i|\mathbf{Z}_i,u_i,\boldsymbol{\theta}) \\ &= \prod_{i=1}^m \phi_q(\mathbf{b}_i|\boldsymbol{\Delta}_i(\mathbf{Z}_i-\mathbf{X}_i^\top\boldsymbol{\beta}),u_i^{-1}\boldsymbol{\Lambda}_i), \end{aligned}$$

where $\boldsymbol{\Delta}_i = \mathbf{D}\mathbf{W}_i^\top\boldsymbol{\Omega}_i^{-1}$, $\boldsymbol{\Lambda}_i = \mathbf{D}-\mathbf{D}\mathbf{W}_i^\top\boldsymbol{\Omega}_i^{-1}\mathbf{W}_i\mathbf{D}$ and $\boldsymbol{\Omega}_i = \mathbf{W}_i\mathbf{D}\mathbf{W}_i^\top+\mathbf{I}_{n_i}$, $i=1,\ldots,m$. To derive the second term on the right-hand side of (12), we use the following result from Chib and Greenberg (1998)

$$\begin{aligned} P(\mathbf{Y}_i &= y_i|\mathbf{b}_i,\mathbf{Z}_i,u_i,\boldsymbol{\theta}) = \mathbb{I}_{(\mathbf{Z}_i\in\mathbb{B}_i)} \\ &= \prod_{j=1}^{n_i}\{\mathbb{I}_{(Z_{ij}>0)}\mathbb{I}_{(Y_{ij}=1)}+\mathbb{I}_{(Z_{ij}\leq 0)}\mathbb{I}_{(Y_{ij}=0)}\}, \end{aligned} \tag{13}$$

which indicates that given $\mathbf{Z}_i$, the conditional probability of $\mathbf{Y}_i$ is independent of $\mathbf{b}_i$ and u_i. Hence, expression (13) implies $P(\mathbf{Y}_i=\mathbf{y}_i|\mathbf{Z}_i,\boldsymbol{\theta}) = \mathbb{I}_{(\mathbf{Z}_i\in\mathbb{B}_i)}$. Since the conditional probability $Z_i|u_i,\boldsymbol{\theta}$ is normally distributed and $U_i\sim \text{Gamma}(v/2,v/2)$, the marginal distribution of $\mathbf{Z}_i|\boldsymbol{\theta}$ follows $t_{n_i}(\mathbf{X}_i\boldsymbol{\beta},\boldsymbol{\Omega}_i,\nu)$. Furthermore, from

$$\begin{aligned} f(\mathbf{Z}_i|\mathbf{Y}_i,\boldsymbol{\theta}) &\propto f(\mathbf{Z}_i,\mathbf{Y}_i|\boldsymbol{\theta}) \\ &= f(\mathbf{Z}_i|\boldsymbol{\theta})P(\mathbf{Y}_i=\mathbf{y}_i|\mathbf{Z}_i,\boldsymbol{\theta}) \\ &= t_{n_i}(\mathbf{Z}_i|X_i^\top\boldsymbol{\beta},\boldsymbol{\Omega}_i,\nu)\mathbb{I}_{(\mathbf{Z}_i\in\mathbb{B}_i)} \end{aligned}$$

we obtain

$$\begin{aligned} f(\mathbf{Z}|\mathbf{Y}_{obs},\boldsymbol{\theta}) &= \prod_{j=1}^{m} f(\mathbf{Z}_i,\mathbf{Y}_i|\boldsymbol{\theta}) \\ &= Tt_{n_i}(\mathbf{Z}_i|X_i^\top\boldsymbol{\beta},\boldsymbol{\Omega}_i,\nu,\mathbb{B}_i). \end{aligned}$$

Using the prior results and the property that, if $\mathbf{Z}|\boldsymbol{\theta}$ follows $t_p(\boldsymbol{\mu},\boldsymbol{\Sigma},\nu)$ and $U \sim \text{Gamma}(\nu/2,\nu/2)$, we have $E[U|\mathbf{Z}] = \dfrac{\nu+p}{\nu+\delta}$ (Lachos et al., 2011), where δ represents the Mahalanobis distance. It follows that:

$$\begin{aligned} E[U_i|\mathbf{Y}_i] &= E\left[E[U_i|\mathbf{Y}_i,\mathbf{Z}_i,\boldsymbol{\theta}]|\mathbf{Y}_i,\boldsymbol{\theta}\right] \\ &= E\left[\frac{\nu+n_i}{\nu+\delta_i}|\mathbf{Y}_i,\boldsymbol{\theta}\right] = \bar{\mathbf{Z}}_i^0, \\ E[U_i\mathbf{Z}_i|\mathbf{Y}_i] &= E\left[\mathbf{Z}_iE[U_i|\mathbf{Y}_i,\mathbf{Z}_i,\boldsymbol{\theta}]|\mathbf{Y}_i,\boldsymbol{\theta}\right] \\ &= E\left[\left(\frac{\nu+n_i}{\nu+\delta_i}\right)Z_i|\mathbf{Y}_i,\boldsymbol{\theta}\right] = \bar{\mathbf{Z}}_i^1, \\ E[U_i\mathbf{b}_i|\mathbf{Y}_i] &= E\left[E\left[U_iE[\mathbf{b}_i|\mathbf{Y}_i,\mathbf{Z}_i,U_i,\boldsymbol{\theta}]|\mathbf{Y}_i,\mathbf{Z}_i,\boldsymbol{\theta}\right]|\mathbf{Y}_i,\boldsymbol{\theta}\right] \\ &= \boldsymbol{\Delta}_i(\bar{\mathbf{Z}}_i^1 - \bar{\mathbf{Z}}_i^0\mathbf{X}_i\boldsymbol{\beta}), \\ E[U_i\mathbf{b}_i\mathbf{b}_i^\top|\mathbf{Y}_i] &= E\left[E\left[U_iE[\mathbf{b}_i\mathbf{b}_i^\top|\mathbf{Y}_i,\mathbf{Z}_i,U_i,\boldsymbol{\theta}]|\mathbf{Y}_i,\mathbf{Z}_i,\boldsymbol{\theta}\right]|\right. \\ &\quad \left.\mathbf{Y}_i,\boldsymbol{\theta}\right] \\ &= \boldsymbol{\Lambda}_i + \boldsymbol{\Delta}_i(\bar{\mathbf{Z}}_i^2 + \boldsymbol{\gamma}_i\boldsymbol{\gamma}_i^\top\bar{\mathbf{Z}}_i^0 - \bar{\mathbf{Z}}_i^1\boldsymbol{\gamma}_i^\top - \boldsymbol{\gamma}_i\bar{\mathbf{Z}}_i^{1^\top})\boldsymbol{\Delta}_i, \\ \mathbf{Z}_i|\mathbf{Y}_i &\sim t_{n_i}(\boldsymbol{\gamma}_i,\boldsymbol{\Omega}_i,\nu)\mathbb{I}_{\mathbb{B}_i}(\mathbf{Z}_i), \end{aligned}$$

where $\bar{\mathbf{Z}}_i^2 = E\left[\dfrac{\nu+n_i}{\nu+\delta_i}\mathbf{Z}_i\mathbf{Z}_i^\top|\mathbf{Y}_i\right]$, $\delta_i = (\mathbf{Z}_i-\boldsymbol{\gamma}_i)^\top\boldsymbol{\Omega}_i^{-1}(\mathbf{Z}_i-\boldsymbol{\gamma}_i)$, $\boldsymbol{\Delta}_i = \mathbf{D}\mathbf{W}_i^\top\boldsymbol{\Omega}_i^{-1}$, $\boldsymbol{\Lambda}_i = \mathbf{D}-\mathbf{D}\mathbf{W}_i^\top\boldsymbol{\Omega}_i^{-1}\mathbf{W}_i\mathbf{D}$, $\boldsymbol{\Omega}_i = \mathbf{W}_i\mathbf{D}\mathbf{W}_i^\top+\mathbf{I}_{n_i}$, $\boldsymbol{\gamma}_i = \mathbf{X}_i\boldsymbol{\beta}$, and $\mathbb{B}_i = B_{i1}\times\ldots,\times B_{in_i}$, where B_{ij} is the interval $(0,\infty)$ if $y_{ij}=1$ and the interval $(-\infty,0]$ if $y_{ij}=0$.

References

Albert, J., Chib, S., 1993. Bayesian analysis of binary and polychotomous response data. Journal of the American Statistical Association 88, 669–679.

Breslow, N., Clayton, D., 1993. Approximate inference in generalized linear mixed models. Journal of the American Statistical Association 88, 9–25.

Chib, S., Greenberg, E., 1998. Analysis of multivariate probit models. Biometrika 85, 347–361.

Czado, C., Santner, T., 1992. The effect of link misspecification on binary regression inference. Journal of Statistical Planning and Inference 33, 213–231.

Delyon, B., Lavielle, M., Moulines, E., 1999. Convergence of a stochastic approximation version of the EM algorithm. Annals of Statistics 27, 94–128.

Fernandez, C., Steel, M. F., 1999. Multivariate student-t regression models: Pitfalls and inference. Biometrika 86, 153–167.

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.

Hjsgaard, S., Halekoh, U., Yan, J., 2005. The r package geepack for generalized estimating equations. Journal of Statistical Software 15, 1–11.

Ho, H. J., Lin, T. I., Chen, H. Y., Wang, W. L., 2012. Some results on the truncated multivariate t distribution. Journal of Statistical Planning and Inference 142, 25–40.

Jamshidian, M., 1999. Adaptive robust regression by using a nonlinear regression program. Journal of Statistical Software 4, 1–25.

Johnson, S., Narasimhan, B., 2011. Package cubature. R package version 1.1-1, URL http://cran.r-project.org/web/packages/cubature/index.html.

Lachos, V. H., Angolini, T., Abanto-Valle, C. A., 2011. On estimation and local influence analysis for measurement errors models under heavy-tailed distributions. Statistical Papers 52, 567–590.

Lange, K. L., Sinsheimer, J. S., 1993. Normal/independent distributions and their applications in robust regression. Journal of Computational and Graphical Statistics 2, 175–198.

Lee, Y., Nelder, J., 2006. Double hierarchical generalized linear models. Applied Statistics 55 (Part 2), 139–185.

Liu, C., 2004. Robit regression: a simple robust alternative to logistic and probit regression. Applied Bayesian modeling and causal inference from incomplete-data perspectives, 227–238.

Lucas, A., 1997. Robustness of the student t basedM-estimator. Communications in Statistics: Theory and Methods 26, 1165–1182.

Matos, L. A., Prates, M. O., H-Chen, M., Lachos, V., 2013. Likelihood-based inference for mixed-effects models with censored response using the multivariate-t distribution. Statistica Sinica 23, 1323–1342.

McCulloch, C., 1994. Maximum likelihood variance components estimation for binary data. Journal of the American Statistical Association 89, 330–335.

McCulloch, C. E., 1997. Maximum Likelihood Algorithms for Generalized Linear Mixed Models. Journal of the American statistical Association 92, 162–170.

McLachlan, G., Krishnan, T., 1997. The EM algorithm and extensions. Wiley New York.

Meng, X., van Dyk, D., 1998. Fast EM-type implementations for mixed effects models. Journal of the Royal Statistical Society, Series B 60, 559–578.

Meza, C., Jaffrzic, F., Foulley, J., 2009. Estimation in the probit normal model for binary outcomes using the SAEM algorithm. Computational Statistics & Data Analysis 53, 1350–1360.

Pinheiro, J. C., Liu, C. H., Wu, Y. N., 2001. Efficient algorithms for robust estimation in linear mixed-effects models using a multivariate t-distribution. Journal of Computational and Graphical Statistics 10, 249–276.

R Core Team, 2013. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
URL `http://www.R-project.org`

Robert, C., Casella, G., Robert, C., 1999. Monte Carlo Statistical Methods. Vol. 2. Springer New York.

Tan, M., Tian, G., Fang, H., 2007. An efficient MCEM algorithm for fitting generalized linear mixed models for correlated binary data. Journal of Statistical Computation and Simulation 77, 929–943.

ApêndiceB

Estimation Methods for Multivariate Tobit Confirmatory Factor Analysis

Estimation Methods for Multivariate Tobit Confirmatory Factor Analysis

Submitted manuscript - Computational Statistics and Data Analysis

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 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 $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Omega}\boldsymbol{\Lambda}^\top + \boldsymbol{\Psi}$. 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 $\mathbf{\Sigma}_i$ can be partitioned as

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

where $vec(.)$ denotes the function which stacks vectors or matrices with the same number of columns. Then we have $\mathbf{y}_i^o \sim N_{p^o}(\mathbf{X}_i^o\boldsymbol{\beta}, \mathbf{\Sigma}_i^{oo})$, $\mathbf{y}_i^c|\mathbf{y}_i^o \sim N_{p^c}(u_i, \mathbf{S}_i)$, where $\boldsymbol{u_i} = \mathbf{X}_i^c\boldsymbol{\beta} + \mathbf{\Sigma}_i^{co}(\mathbf{\Sigma}_i^{oo})^{-1}(\mathbf{y}_i^o - \mathbf{X}_i^o\boldsymbol{\beta})$ and $\mathbf{S}_i = \mathbf{\Sigma}_i^{cc} - \mathbf{\Sigma}_i^{co}(\mathbf{\Sigma}_i^{oo})^{-1}\mathbf{\Sigma}_i^{oc}$. 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

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

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

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

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 $\mathbf{A}_\beta = (\mathbf{S}_\beta^{-1}+\sum_{i=1}^n \mathbf{X}_i^\top \mathbf{\Psi}^{-1}\mathbf{X}_i)^{-1}$, $\boldsymbol{u_\beta = (\beta_0^\top S_\beta^{-1}+\sum_{i=1}^n (y_i - X_i^\top \beta_0)^\top \Psi^{-1} X_i^\top)}$, $s = \sum_{i=1}^n (\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \Lambda_r^\top \mathbf{z}_i)$, $\mathbf{A}_\Lambda = (\sum_{i=1}^n \mathbf{z}_i\mathbf{z}_i^\top \mathbf{\Psi}^{-1}+\mathbf{\Lambda}_0^{-1})^{-1}$, $\boldsymbol{u_\Lambda = \sum_{i=1}^n A_{ir}^\top \Psi^{-1} z_{ir} + \Lambda_0^\top S_\Lambda^{-1}}$, $\mathbf{A}_{ir} = \mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \sum_{l\neq r}^{p_{l=1}} \Lambda_l \mathbf{z}_{il}$, $r = 1, ..., q$.

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.

		EM results				Bayesian results			
Censoring	Parameters	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.

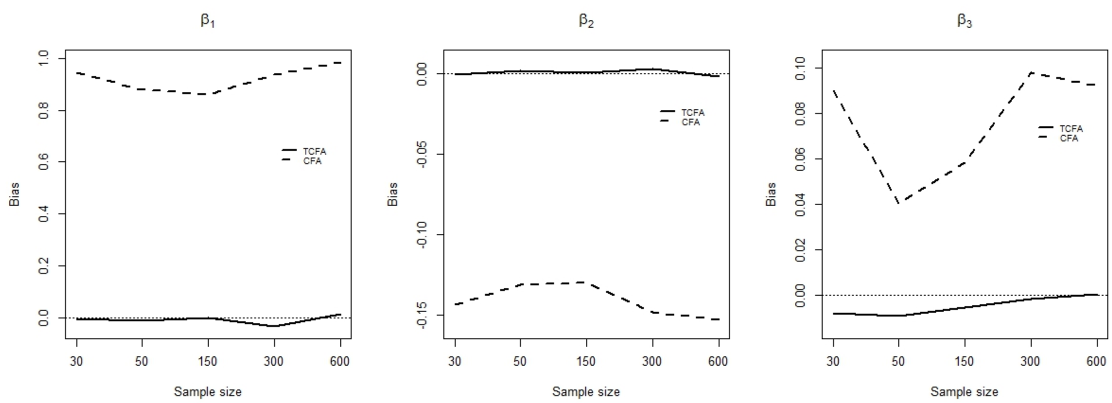

Figure 1: Simulated data. Bias from the parameter $\boldsymbol{\beta}$.

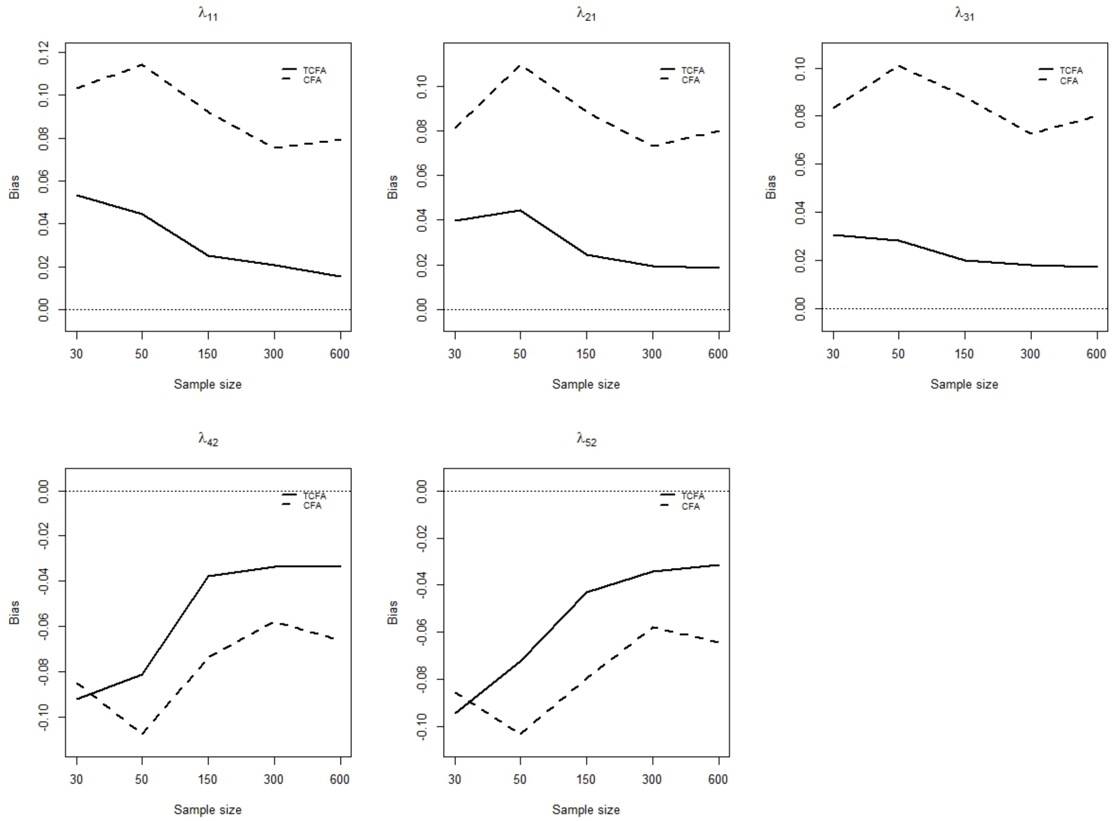

Figure 2: Simulated data. Bias from the parameter $\mathbf{\Lambda}$.

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 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

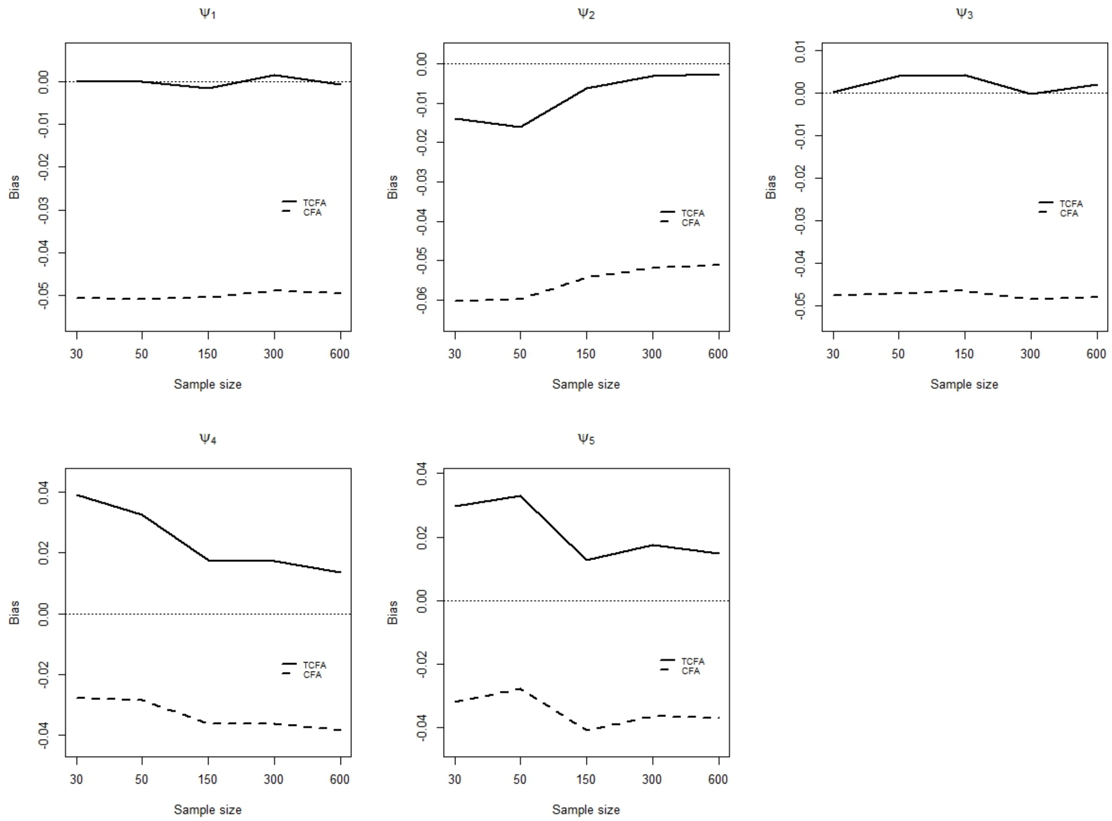

Figure 3: Simulated data. Bias from the parameter $\mathbf{\Psi}$.

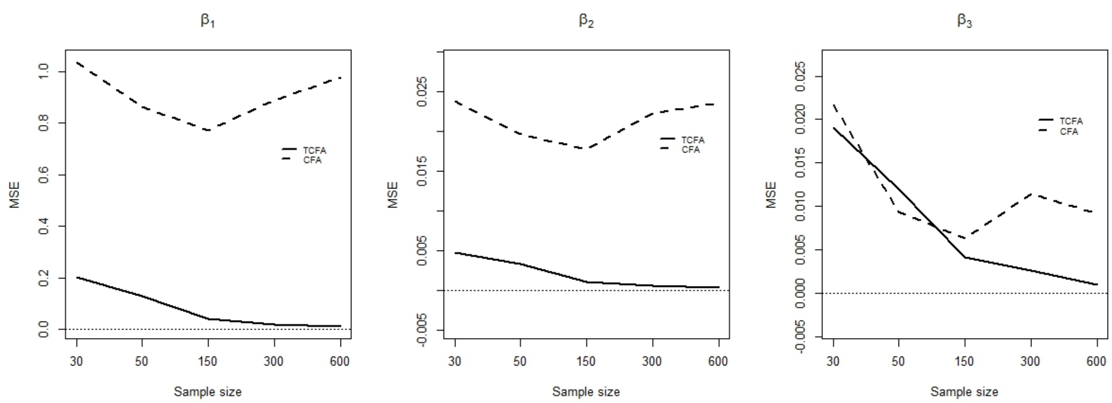

Figure 4: Simulated data. MSE from the parameter $\boldsymbol{\beta}$.

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).

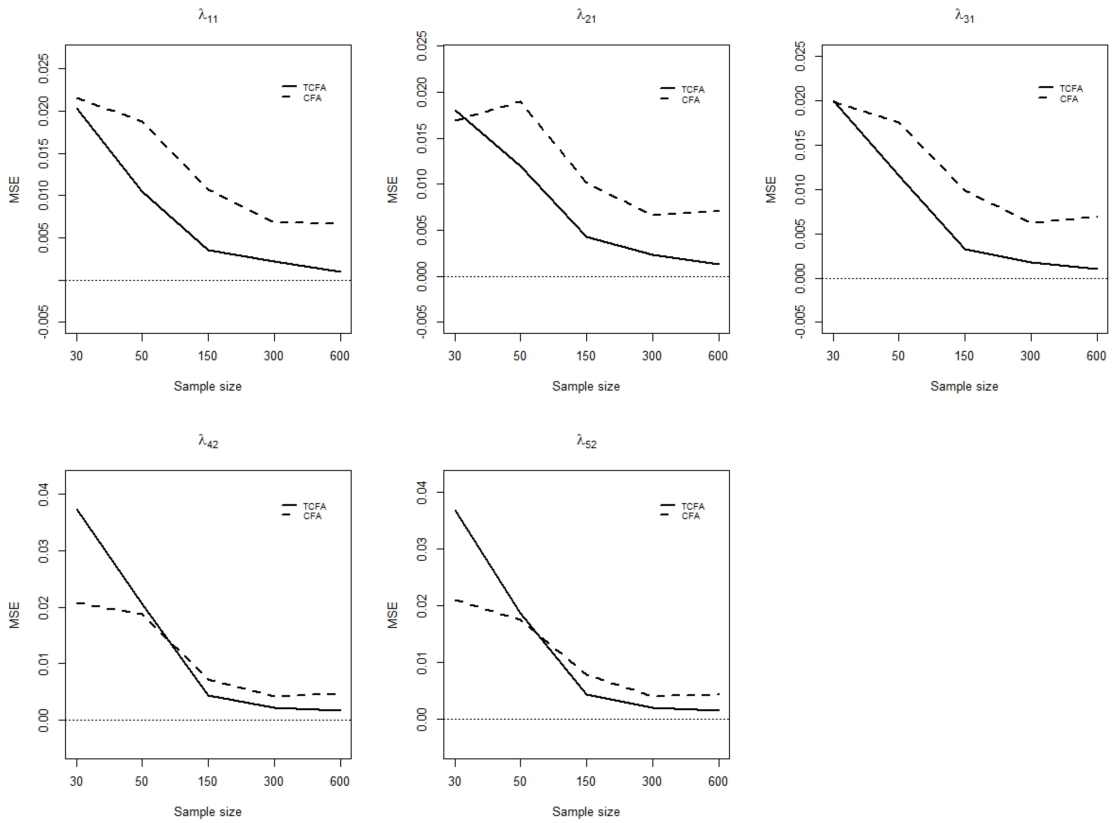

Figure 5: Simulated data. MSE from the parameter $\mathbf{\Lambda}$.

This transformation indicates that students with high score in fluency in Spanish 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

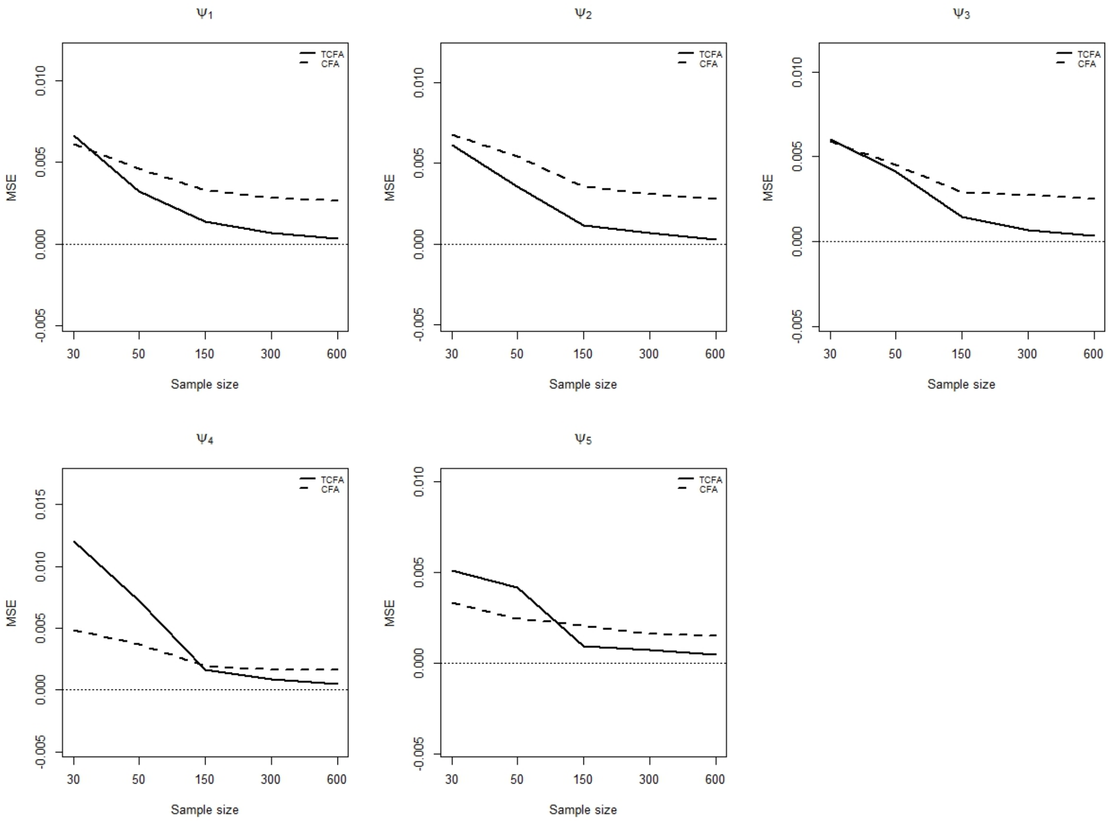

Figure 6: Simulated data. MSE from the parameter $\mathbf{\Psi}$.

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 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), $\mathbf{\Lambda}$ and $\mathbf{\Psi}$, 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-

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

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: 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		

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 $\mathbf{\Psi}$) are relatively 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 $J-$distance. Table 4 shows the characteristics of these students in the dataset.

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 and 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 and 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

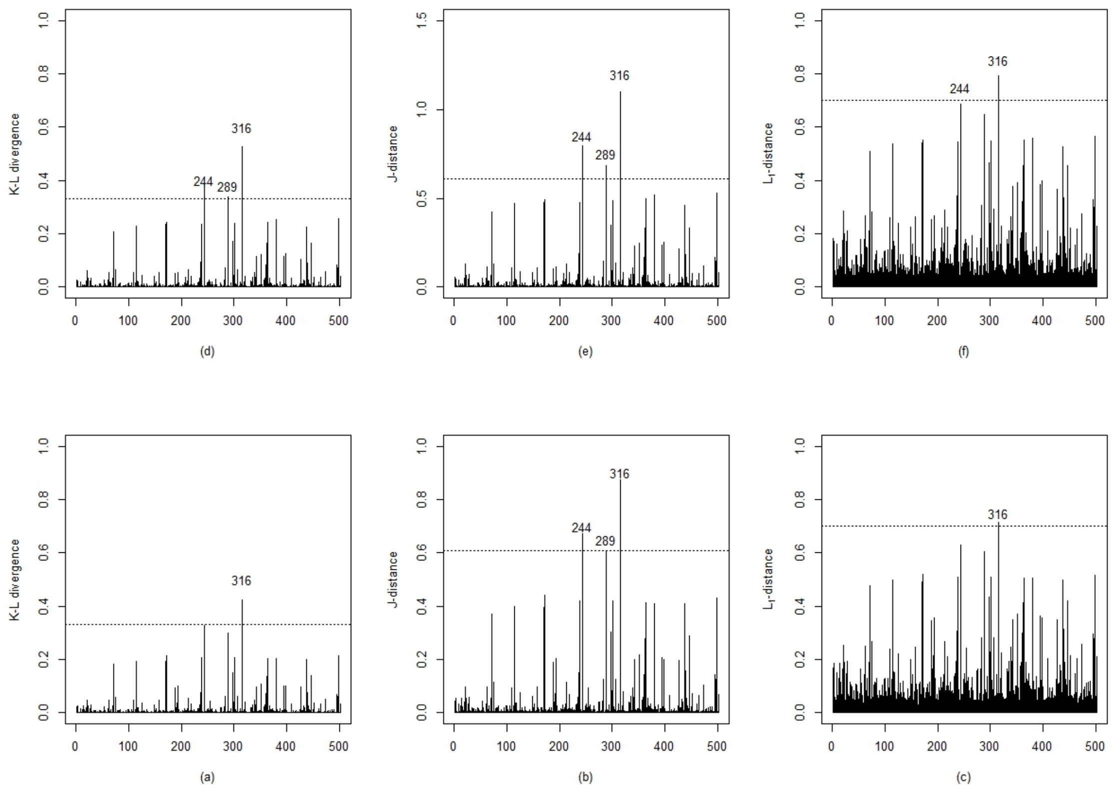

Figure 7: Index plots of d_q measure for the EGRA data. The first three plots (a-c) refer to the CFA model and the last three (d-f) to the TCFA.

data. The three observations that stood out according to the analysis of influence diagnostic measures are highlighted.

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).

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

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

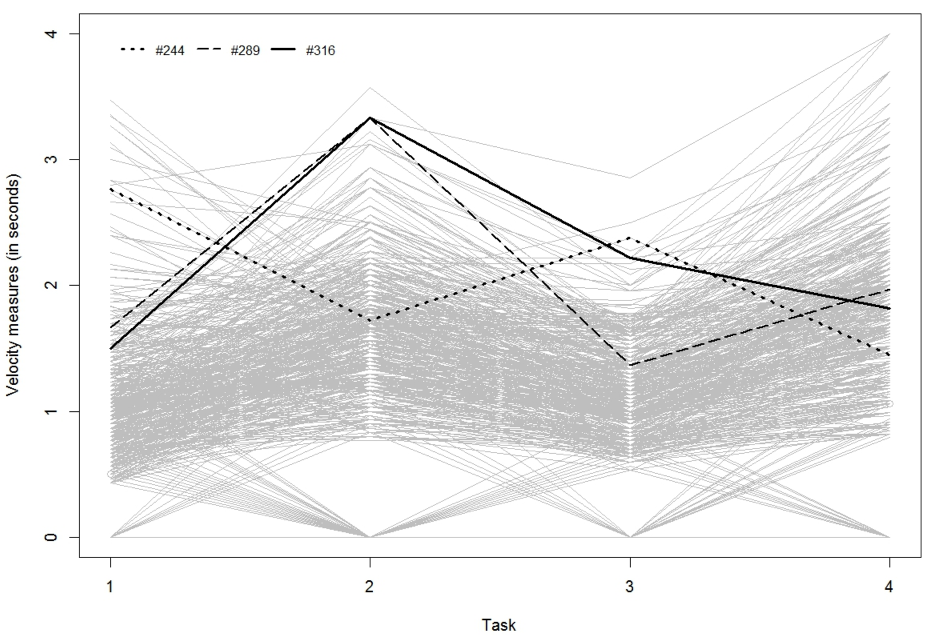

Figure 8: EGRA data. Individual velocity measures for each task. The trajectories for the influential individuals are numbered.

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

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

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.

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.

Apêndice C

Likelihood-Based Inference for Tobit Confirmatory Factor Analysis Using the Multivariate t-Distribution

Likelihood-Based Inference for Tobit Confirmatory Factor Analysis Using the Multivariate Student-t Distribution

Submitted manuscript - Statistics and Computing

Abstract

Factor analysis models have been one of the most popular multivariate methods for data analysis among psychometricians and behavioral researchers. These models, originally developed for normally distributed observed variables, can be seriously affected by the presence of influential observations and censored data. Motivated by this situation, in this paper we propose a likelihood-based estimation for a multivariate tobit confirmatory factor analysis model using the Student-t distribution (t-TCFA model). An EM-type algorithm is developed for computing the maximum likelihood estimates, obtaining as a byproduct the standard errors of the fixed effects and the exact likelihood value. Unlike other approaches proposed in the literature, our exact EM-type algorithm uses closed form expressions at the E-step based on the first two moments of a truncated multivariate Student-t distribution with the advantage that these expressions can be computed using standard statistical software. The performance of the proposed method is illustrated through a simulation study and a real dataset of Early Grade Reading Assessment (EGRA) test scores.

Keywords: Factor analysis; Latent variable; Multivariate tobit model; Student-t distribution.

1 Introduction

Confirmatory and exploratory factor analysis (CFA and EFA) are popular multivariate techniques used for the analysis of interdependencies among observed variables and underlying theoretical constructs or latent variables, called factors. The main difference between CFA and EFA is that while in CFA the researcher has to specify the number of factors, the correlations between these factors and the variables that load the factors previously, the EFA is used when there is no information about the latent variables and the researcher has to identify the underlying relationship between the measured variables.

Originally, CFA and EFA models were developed for continuous and normally distributed observed variables because of their mathematical tractability. However, when a departure from normality is observed, the normality assumption can generate biased and misleading inferences. In order to overcome this problem, some robustness strategies based on the Student-t distribution are proposed in the statistical literature related principally to CFA. For example, Zhang et al. (2013) introduce the Student-t factor analysis (t-CFA) and propose an EM-type algorithm for maximum likelihood (ML) estimation. They demonstrate the robustness for the parameter estimation in t-CFA through extensive simulations. McLachlan et al. (2007) and Wang and Lin (2013) propose finite mixtures of factor analyzers based on the Student-t distribution. Lin et al. (2013) advocate the use of the skew-t distribution (Azzalini and Genton, 2008) for robust estimation in the context of CFA models.

In some studies, the measures can be subjected to some upper and lower thresholds below or above which they are not quantifiable. For example, in education measurement, there are some tests in which the students have to carry out some tasks within a maximum period of time. As a consequence of this type of evaluation, the responses (or specifically the scores) are censored because of the time limit. As was noted by Muthén (1989) and Kamakura and Wedel (2001), to ignore the censoring scheme in this case may produce biased estimates of the factor pattern. Thus. many solutions have been proposed to deal with the problem of censored responses in factor analysis. Muthén (1989) proposed a CFA model for censored data developing a three-stage estimation procedure and Kamakura and Wedel (2001) extended that work considering a framework for both CFA and EFA modeling for censored data. Zhou and Liu (2009) proposed a Monte Carlo EM (MCEM) algorithm for the estimation of the CFA under the assumption that the latent variable follows a multivariate Tobit model and Zhou and Tan (2010) extended that work by considering a Student-t Tobit model. Alternatively, Bayesian inference can be carried out with Markov chain Monte Carlo (MCMC) and implemented via Gibbs sampling (see, for example Costa et al., 2013). However, by their nature MCEM and MCMC methods are expensive propositions, due to a combination of Monte Carlo simulation with iterative procedures.

In this paper, we propose a robust CFA model based on the multivariate Student-t distribution for censored data so that the t-TCFA model is defined and a fully likelihood-based approach is carried out, including the implementation of an exact EM-type algorithm for the ML estimation. Like Matos et al. (2013b) (see also Prates et al., 2013), we show that the E-step reduces to computing the first two moments of a truncated multivariate Student-t distribution. The likelihood function is easily computed as a byproduct of the E-step and is used for monitoring convergence and for model selection using the Akaike information criterion (AIC) (Akaike, 1974), the Bayesian information criterion (BIC) (Schwarz, 1978) or the consistent Akaike information criterion (CAIc) (Bozdogan, 1987). Thus, our proposal extends the work of Zhou and Tan (2010) in two ways. First, we consider a general censoring scheme and second we provide an exact EM-type algorithm for robust estimation in t-TCFA models.

The plan of the paper is as follows. In Section 2, we introduce some basic notation and an outline of the main results related to the multivariate and truncated Student-t distributions. After that, the multivariate tobit confirmatory factor analysis model under the Student-t distribution is defined in Section 3. Section 4 describes the steps of the EM-type algorithm for obtaining the ML estimates. In Section 5 we apply our proposed method to a real educational data set corresponding to the Early Grade Reading Assessment (EGRA) test scores from Peruvian students. Finally, some concluding remarks are provided in Section 6.

2 Preliminaries

In this section we provide some useful results related to the multivariate Student-t distribution. We start with the definition of a Student-t random vector.

A random vector $\mathbf{Y}$ is said to follow a p-variate Student-t distribution with location vector $\boldsymbol{\mu}$, scale matrix $\boldsymbol{\Sigma}$ and degrees of freedom ν, denoted by $t_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$, if it can be represented by

$$\mathbf{Y} = \boldsymbol{\mu} + U^{-1/2}\mathbf{Z}, \tag{1}$$

where $\mathbf{Z} \sim N_p(\mathbf{0}, \boldsymbol{\Sigma})$ and $U \sim \text{Gamma}(\nu/2, \nu/2)$ are independent random objects and $\text{Gamma}(a, b)$ stands for a gamma distribution with mean a/b and density denoted by $G(\cdot|a, b)$. Then, we obtain the probability density function (*pdf*) of $\mathbf{Y}$, given by

$$t_p(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu) = \frac{\Gamma(\frac{p+\nu}{2})\nu^{-p/2}}{\Gamma(\frac{\nu}{2})\pi^{p/2}}|\boldsymbol{\Sigma}|^{-1/2}\left(1 + \frac{\delta}{\nu}\right)^{-\frac{(p+\nu)}{2}},$$

where $\Gamma(\cdot)$ is the standard gamma function and $\delta = (\mathbf{y} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{y} - \boldsymbol{\mu})$ is the Mahalanobis distance. The cumulative distribution function (*cdf*) of the Student-t distribution is denoted by $T_p(\cdot|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$. If $\nu > 1$, then $\boldsymbol{\mu}$ is the mean of $\mathbf{Y}$, and if $\nu > 2$ then $\nu(\nu-2)^{-1}\boldsymbol{\Sigma}$ is its covariance matrix. As ν tends to infinity, U converges almost surely to one and, marginally, $\mathbf{Y}$ converges in distribution to a multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.

The following result, taken from Matos et al. (2013b), presents the marginal-conditional decomposition of the multivariate Student-t distribution.

Proposition 1. *Let $\mathbf{Y} \sim t_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$. If $\mathbf{Y}$ is partitioned as $\mathbf{Y} = (\mathbf{Y}_1^\top, \mathbf{Y}_2^\top)^\top$, with $dim(\mathbf{Y}_1^\top) = p_1$ and $dim(\mathbf{Y}_2^\top) = p_2$ where $p = p_1 + p_2$, and $\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}$ and $\boldsymbol{\mu} = (\boldsymbol{\mu}_1^\top, \boldsymbol{\mu}_2^\top)^\top$ are the corresponding partitions of $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}$, then:*

- $\mathbf{Y}_1^\top \sim t_{p_1}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}, \nu)$.
- *The conditional cdf of $\mathbf{Y}_2^\top | \mathbf{Y}_1^\top = \mathbf{y}_1$ is given by*

$$P(\mathbf{Y}_2^\top \leq \mathbf{y}_2 | \mathbf{Y}_1^\top = \mathbf{y}_1) = T_{p_2}\left(\mathbf{y}_2 | \boldsymbol{\mu}_{2.1}, \hat{\boldsymbol{\Sigma}}_{22.1}, \nu + p_1\right),$$

i.e., $\mathbf{Y}_2^\top | \mathbf{Y}_1^\top = \mathbf{y}_1 \sim t_{p_2}\left(\boldsymbol{\mu}_{2.1}, \hat{\boldsymbol{\Sigma}}_{22.1}, \nu + p_1\right)$, where $\hat{\boldsymbol{\Sigma}}_{22.1} = \left(\frac{\nu+\delta_1}{\nu+p_1}\right)\boldsymbol{\Sigma}_{22.1}$, $\delta_1 = (\mathbf{y}_1 - \boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}_{11}^{-1}(\mathbf{y}_1 - \boldsymbol{\mu}_1)$, $\boldsymbol{\Sigma}_{22.1} = \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}$, $\boldsymbol{\mu}_{2.1} = \boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\mathbf{y}_1 - \boldsymbol{\mu}_1)$.

Now, $Tt_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu; \mathbb{A})$ denotes a p-variate truncated Student-t distribution whose *pdf* is defined as the density $t_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$ lying within a right-truncated hyperplane

$$\mathbf{A} = \{\mathbf{x} = (x_1, \ldots, x_p)^\top | x_1 \leq a_1, \ldots, x_p \leq a_p\}. \tag{2}$$

Specifically, we say that the p-dimensional vector $\mathbf{X} \sim Tt_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu; \mathbb{A})$ if its *pdf* is given by $f(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu; \mathbb{A}) = \dfrac{t_p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)}{T_p(\mathbf{a}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)}\mathbb{I}_A(\mathbf{x})$, where $\mathbf{a} = (a_1, \ldots, a_p)^\top$ and $\mathbb{I}_A(\mathbf{x})$ is the indicator function, whose value is equal to one if $\mathbf{x} \in \mathbb{A}$ and zero otherwise. The following results provide the moments of the right-truncated multivariate Student-t distribution. These results are useful in the implementation of the EM algorithm for our t-TCFA model. Their proofs are given in Appendix A.

Proposition 2. *If* $\mathbf{X} \sim Tt_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu; \mathbb{A})$ *with* $\mathbb{A}$ *as defined in (2), then the kth moment of* $\mathbf{X}$ *for* $k = 0, 1, 2$*, is given by*

$$E\left[\left(\frac{\nu+p}{\nu+\delta}\right)^r \mathbf{X}^{(k)}\right] = c_p(\nu, r) \times \frac{T_p(\mathbf{a}|\boldsymbol{\mu}, \boldsymbol{\Sigma}^*, \nu+2r)}{T_p(\mathbf{a}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)} E_{\mathbf{W}}\left[\mathbf{W}^{(k)}\right],$$

$$\mathbf{W} \sim Tt_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}^*, \nu+2r; \mathbb{A}),$$

where $c_p(\nu, r) = \left(\frac{\nu+p}{\nu}\right)^r \left(\frac{\Gamma((p+\nu)/2)\Gamma((\nu+2r)/2)}{\Gamma(\nu/2)\Gamma((p+\nu+2r)/2)}\right)$, $\delta = (\mathbf{X}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{X}-\boldsymbol{\mu})$, $\mathbf{a} = (a_1, \ldots, a_p)^\top$, $\boldsymbol{\Sigma}^* = \frac{\nu}{\nu+2r}\boldsymbol{\Sigma}$, $\mathbf{X}^{(0)} = 1$, $\mathbf{X}^{(1)} = \mathbf{X}$, $\mathbf{X}^{(2)} = \mathbf{X}\mathbf{X}^\top$ *and* $\nu + 2r > 0$.

Proposition 3. *Let* $\mathbf{X} \sim Tt_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu; \mathbb{A})$ *with* $\mathbb{A}$ *as defined in (2). Consider the partition* $\mathbf{X} = (\mathbf{X}_1^\top, \mathbf{X}_2^\top)^\top$*, with dim*$(\mathbf{X}_1^\top) = p_1$*, dim*$(\mathbf{X}_2^\top) = p_2$, $p = p_1 + p_2$, $\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}$*, and* $\boldsymbol{\mu} = (\boldsymbol{\mu}_1^\top, \boldsymbol{\mu}_2^\top)^\top$ $\mathbf{z} = (\mathbf{z}^{x_1}, \mathbf{z}^{x_2})$ *and* $\mathbb{A} = (\mathbb{A}^{x_1}, \mathbb{A}^{x_2})$*. Then, under the notation given in Proposition 2, we have:*

$$E\left[\left(\frac{\nu+p}{\nu+\delta}\right)^r \mathbf{X}_2^{(k)}|\mathbf{X}_1\right] = \frac{d_p(p_1, \nu, r)}{(\nu+\delta_1)^r} \times \frac{T_{p_2}(\mathbf{a}^{\mathbf{x_2}}|\boldsymbol{\mu}_{2.1}, \hat{\boldsymbol{\Sigma}}^*_{22.1}, \nu+p_1+2r)}{T_{p_2}(\mathbf{a}^{\mathbf{x_2}}|\boldsymbol{\mu}_{2.1}, \hat{\boldsymbol{\Sigma}}^*_{22.1}, \nu+p_1)} E_{\mathbf{W}}\left[\mathbf{W}^{(k)}\right],$$

$$\mathbf{W} \sim Tt_{p_2}(\boldsymbol{\mu}_{2.1}, \hat{\boldsymbol{\Sigma}}^*_{22.1}, \nu+p_1+2r; \mathbb{A}^{x_2}),$$

where $d_p(p_1, \nu, r) = (\nu+p)^r \left(\frac{\Gamma((p+\nu)/2)\Gamma((p_1+\nu+2r)/2)}{\Gamma((p_1+\nu)/2)\Gamma((p+\nu+2r)/2)}\right)$, $\delta = (\mathbf{X}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{X}-\boldsymbol{\mu})$, $\delta_1 = (\mathbf{X}_1 - \boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}_{11}^{-1}(\mathbf{X}_1 - \boldsymbol{\mu}_1)$, $\mathbf{a}_{x_2} = (a_1, \ldots, a_{p_2})^\top$, $\boldsymbol{\Sigma}^*_{22.1} = \frac{\nu+\delta_1}{\nu+2r+p_1}\boldsymbol{\Sigma}_{22.1}$, $\mathbf{X}^{(0)} = 1$, $\mathbf{X}^{(1)} = \mathbf{X}$, $\mathbf{X}^{(2)} = \mathbf{X}\mathbf{X}^\top$, $\nu + p_1 + 2r > 0$ *and* $k = 0, 1, 2$.

Expressions for $E[\mathbf{W}]$ and $E[\mathbf{W}\mathbf{W}^\top]$ where $\mathbf{W} \sim Tt_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu; \mathbb{A})$ have been recently developed in closed form by Ho et al. (2012). These expressions depend on the multivariate Student-t *cdf*. They can be computated by using functions for the numerical evaluation of the Student-t *cdf* such as *pmvt*() function of the *mvtnorm* package (Genz et al., 2008) from `R`.

3 The model

3.1 The Gaussian tobit confirmatory factor analysis model

The Gaussian (or normal) multivariate tobit CFA model (TCFA) with covariates described by Zhou and Liu (2009), is specified as follows:

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\Lambda}\mathbf{z}_i + \boldsymbol{\epsilon}_i, \tag{3}$$

where $\mathbf{z}_i \overset{iid}{\sim} N_q(\mathbf{0}, \boldsymbol{\Omega})$ is independent of $\boldsymbol{\epsilon}_i \overset{ind.}{\sim} N_p(\mathbf{0}, \boldsymbol{\Psi})$, for $i = 1, \ldots, n$. The subscript i denotes 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 $k \times 1$; $\boldsymbol{\beta}$ is the vector of fixed effects 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, the dispersion matrix $\boldsymbol{\Omega}$ is an arbitrary covariance or correlation matrix and $\boldsymbol{\Psi}$ is a diagonal covariance matrix.

In the present formulation, we consider the case where the response y_{ij} is not fully observed for all i, j, $i = 1, \ldots, n$, $j = 1, \ldots, p$. Following Vaida and Liu (2009) and Matos et al. (2013a), let $(\mathbf{V}_i, \mathbf{C}_i)$ be the observed data for the i-th subject, where $\mathbf{V}_i$ represents the vector of uncensored reading or censoring level and $\mathbf{C}_i$ is the vector of censoring indicators with components C_{ij}. Then, we have

$$\begin{aligned} y_{ij} &\leq V_{ij} \quad \text{if} \quad C_{ij} = 1, \\ y_{ij} &= V_{ij} \quad \text{if} \quad C_{ij} = 0. \end{aligned} \tag{4}$$

By combining (3) and (4), the TCFA model is formulated. In this work, we will use a left censoring pattern, but extensions to right or arbitrary censoring are immediate.

As is well known, factor analysis models are not identified because if one considers $\boldsymbol{\Lambda}^* = \boldsymbol{\Lambda}\boldsymbol{\Gamma}^{-1}$ and $\mathbf{z}_i^* = \boldsymbol{\Gamma}\mathbf{z}_i$, for any nonsingular matrix $\boldsymbol{\Gamma}$, the same model (3) is obtained. To solve this identification problem, Zhou and Liu (2009) recommended to fix elements of $\mathbf{\Lambda}$ and/or $\mathbf{\Omega}$. As a confirmatory analysis, we will assume $\mathbf{\Omega}$ as known in this work and, as a consequence, avoid the analysis for not identifiable models.

The classic inference about the parameter vector $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \mathbf{\Lambda}, \mathbf{\Psi})^\top$ is based on the marginal distribution of $\mathbf{y}_i$. For complete data, we have that marginally $\mathbf{y}_i \overset{\text{ind.}}{\sim} N_p(\mathbf{X}_i\boldsymbol{\beta}, \mathbf{\Sigma})$, $i = 1, \ldots, n$, where $\mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Omega}\mathbf{\Lambda}^\top + \mathbf{\Psi}$. For responses a censoring pattern as in (4), we have that $\mathbf{y}_i \sim TN_p(\mathbf{X}_i\boldsymbol{\beta}, \mathbf{\Sigma}_i; \mathbb{A})$, where $TN_p(\cdot; \mathbb{A})$ denotes the truncated normal distribution on the interval $\mathbb{A}$, where $\mathbb{A}_i = A_{i1} \times \ldots \times A_{in_i}$, 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 TCFA model, the first step is to treat separately the observed and censored components of $\mathbf{y}_i$. Let $\mathbf{y}_i^o$ be the p^o vector of observed outcomes 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 $\mathbf{\Sigma}_i$ can be partitioned as:

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

where $vec(\cdot)$ denotes the function which stacks vectors or matrices of the same number of columns. Then, we have $\mathbf{y}_i^o \sim N_{p_i^o}(\mathbf{X}_i^o\boldsymbol{\beta}, \mathbf{\Sigma}_i^{oo})$, $\mathbf{y}_i^c|\mathbf{y}_i^o \sim N_{p_i^c}(\boldsymbol{\mu}_i, \mathbf{S}_i)$, where $\boldsymbol{\mu}_i = \mathbf{X}_i^c\boldsymbol{\beta} + \mathbf{\Sigma}_i^{co}(\mathbf{\Sigma}_i^{oo})^{-1}(\mathbf{y}_i^o - \mathbf{X}_i^o\boldsymbol{\beta})$ and $\mathbf{S}_i = \mathbf{\Sigma}_i^{cc} - \mathbf{\Sigma}_i^{co}(\mathbf{\Sigma}_i^{oo})^{-1}\mathbf{\Sigma}_i^{oc}$.

Now, let $\Phi_p(\mathbf{u}; \mathbf{a}, \mathbf{A})$ and $\phi_p(\mathbf{u}; \mathbf{a}, \mathbf{A})$ be the *cdf* (left tail) and *pdf*, respectively, of an $N_p(\mathbf{a}, \mathbf{A})$ computed at $\mathbf{u}$. From Vaida and Liu (2009) (see also, Jacqmin-Gadda et al., 2000), the likelihood function for cluster i (using conditional probability arguments) is given by

$$\begin{aligned} L_i(\boldsymbol{\theta}) = f(\mathbf{y}_i|\boldsymbol{\theta}) &= P(\mathbf{V}_i|\mathbf{C}_i, \boldsymbol{\theta}) = P(\mathbf{y}_i^c \leq \mathbf{V}_i^c|\mathbf{y}_i^o = \mathbf{V}_i^o, \boldsymbol{\theta})P(\mathbf{y}_i^o = \mathbf{V}_i^o|\boldsymbol{\theta}), \\ &= P(\mathbf{y}_i^c \leq \mathbf{V}_i^c|\mathbf{y}_i^o, \boldsymbol{\theta})f(\mathbf{y}_i^o|\boldsymbol{\theta}) \\ &= \phi_{p_i^o}(\mathbf{y}_i^o; \mathbf{X}_i^o\boldsymbol{\beta}, \mathbf{\Sigma}_i^{oo})\Phi_{p_i^c}(\mathbf{V}_i^c; \boldsymbol{\mu}_i, \mathbf{S}_i), \end{aligned} \tag{5}$$

which can be evaluated without much computational burden through the routine *mvtnorm()* available in `R` (see Genz et al., 2008; R Development Core Team, 2009). The log-likelihood function for the observed data is given by $\ell(\boldsymbol{\theta}|\mathbf{y}) = \sum_{i=1}^n \{\log L_i\}$. The estimates obtained by maximizing the log-likelihood function $\ell(\boldsymbol{\theta}|\mathbf{y})$ are thus the ML estimates for the TCFA model defined in (3) and (4). In the next section, we propose the multivariate tobit factor analysis model by assuming a multivariate Student-t distribution in the random terms (t-TCFA model).

3.2 The Student-t tobit confirmatory factor analysis model

Zhou and Tan (2010) proposed a robust hierarchical tobit factor analysis model in which the random effects and the within-subject errors have multivariate Student-t distributions (t-CFA) and assumed that the mixing random variable U for the two sources of variability in the model has the same parameters, *i.e.* the same degrees of freedom ν. Thus, to obtain robust estimates of the parameters, we proceed as in Zhou and Tan (2010) (see also, Matos et al., 2013b) by considering censored data and a generalization of the Gaussian TCFA model defined in (3) and (4), with

$$(\mathbf{z}_i, \boldsymbol{\epsilon}_i)^\top \sim t_{p+q}\{\mathbf{0}, \mathrm{Diag}(\boldsymbol{\Omega}, \boldsymbol{\Psi}), \nu\}, \; i = 1, \ldots, n, \tag{6}$$

or equivalently by defining $\mathbf{z}_i \overset{iid}{\sim} t_q(\mathbf{0}, \boldsymbol{\Omega}, \nu)$ and $\boldsymbol{\epsilon}_i \overset{iid}{\sim} t_p(\mathbf{0}, \boldsymbol{\Psi}, \nu)$ where $\mathrm{Diag}(\mathbf{A}, \mathbf{B})$ is a block diagonal matrix whose elements are the square matrices $\mathbf{A}$ and $\mathbf{B}$. Note that $\mathbf{z}_i$ and $\boldsymbol{\epsilon}_i$ are uncorrelated since $\mathrm{Cov}(\mathbf{z}_i, \boldsymbol{\epsilon}_i) = E[\mathbf{z}_i \boldsymbol{\epsilon}_i^\top] = E[E(\mathbf{z}_i \boldsymbol{\epsilon}_i^\top | U_i)] = \mathbf{0}$, where U_i is a scalar generated from $\mathrm{Gamma}(\nu/2, \nu/2)$.

Like the Gaussian case, classic inference on the parameter vector $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \boldsymbol{\Lambda}, \boldsymbol{\Psi})^\top$ is based on the marginal distribution of $\mathbf{y}_i$, which, for the complete data (t-CFA), is $\mathbf{y}_i \overset{\text{ind.}}{\sim} t_p(\mathbf{X}_i \boldsymbol{\beta}, \boldsymbol{\Sigma}, \nu)$, $i = 1, \ldots, n$, where $\boldsymbol{\Sigma} = \boldsymbol{\Lambda} \boldsymbol{\Omega} \boldsymbol{\Lambda}^\top + \boldsymbol{\Psi}$. The estimates from the multivariate t-CFA are more robust against outliers than those based on the standard CFA. In a simulation study, Zhang et al. (2013) showed that the t-CFA substantially outperforms the normal or standard CFA when outliers are present in the data. This issue has also been discussed by Wu (2010) and Matos et al. (2013b) in the context of censored mixed-effects models.

As in the censored normal case (Subsection 3.1), the likelihood function can also be computed treating separately the observed and censored components of $\mathbf{y}_i$. Then, under the notation introduced in Section 2, we have that $\boldsymbol{\mu}_i^{co} = \mathbf{X}_i^c \boldsymbol{\beta} + \boldsymbol{\Sigma}_i^{co} (\boldsymbol{\Sigma}_i^{oo})^{-1} (\mathbf{y}_i^o - \mathbf{X}_i^o \boldsymbol{\beta})$ and $\mathbf{S}_i = \left(\frac{\nu + \delta_i^o}{\nu + p_i^o}\right) \boldsymbol{\Sigma}_i^{cc.o}$, with $\boldsymbol{\Sigma}_i^{cc.o} = \boldsymbol{\Sigma}_i^{cc} - \boldsymbol{\Sigma}_i^{co} (\boldsymbol{\Sigma}_i^{oo})^{-1} \boldsymbol{\Sigma}_i^{oc}$ and $\delta_i^o = (\mathbf{y}_i^o - \mathbf{X}_i^o \boldsymbol{\beta})^\top (\boldsymbol{\Sigma}_i^{oo})^{-1} (\mathbf{y}_i^o - \mathbf{X}_i^o \boldsymbol{\beta})$. Thus, the likelihood function for cluster i is given by

$$L_i(\boldsymbol{\theta}) = t_{p_i^o}(\mathbf{y}_i^o; \mathbf{X}_i^o \boldsymbol{\beta}, \boldsymbol{\Sigma}_i^{oo}, \nu) T_{p_i^c}(\mathbf{V}_i^c; \boldsymbol{\mu}_i^{co}, \mathbf{S}_i^{co}, \nu + p_i^o). \tag{7}$$

The log-likelihood function for the observed data ($\ell(\boldsymbol{\theta}|\mathbf{y}) = \sum_{i=1}^n \{\log L_i(\boldsymbol{\theta})\}$) can be computed at each step of the EM-type algorithm without additional computational difficulties, because the L_i functions have already been computed at the E-step. In addition, the log-likelihood function can be used to monitor the convergence of the EM algorithm and for the model selection by AIC $= -2\ell(\boldsymbol{\theta}|\mathbf{y}) + 2T$, BIC $= -2\ell(\boldsymbol{\theta}|\mathbf{y}) + Tln(n)$, or CAIc $= -2\ell(\boldsymbol{\theta}|\mathbf{y}) + T(ln(n) + 1)$, where T denotes the number of parameters in the model and n denotes the sample size.

Lucas (1997) carried out an interesting study of the robust aspects of the Student-t M-estimator in the univariate case using influence functions. He showed that the protection against outliers is preserved only if the degrees of freedom parameter is fixed. In this paper, we consider this assumption using a model selection procedure based on AIC or BIC to choose the most appropriate value of ν (see Lange et al., 1989; Meza et al., 2012). Thus, hereafter we consider that the parameter vector is $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \boldsymbol{\Lambda}, \boldsymbol{\Psi})^\top$.

4 Efficient EM algorithm for ML estimation

The EM algorithm originally proposed by Dempster et al. (1977) has several appealing features such as stability of monotone convergence with each iteration increasing the likelihood and simplicity of implementation. However, ML estimation in model (3)-(4) and (6) is complicated and the EM algorithm is less advisable due to the computational difficulty in the M-step. To cope with this problem, we apply an extension of the EM algorithm, called the ECM algorithm (Meng and Rubin, 1993), which shares the appealing features of the EM and typically has a faster convergence rate than the EM.

Let $\mathbf{y} = (\mathbf{y}_1^\top, \ldots, \mathbf{y}_n^\top)^\top$, $\mathbf{z} = (\mathbf{z}_1^\top, \ldots, \mathbf{z}_n^\top)^\top$, $\mathbf{V} = vec(\mathbf{V}_1, \ldots, \mathbf{V}_n)$ and $\mathbf{C} = vec(\mathbf{C}_1, \ldots, \mathbf{C}_n)$, such that we observe $(\mathbf{V}_i, \mathbf{C}_i)$ for the ith subject. In their estimation procedure, $\mathbf{z}$, $\mathbf{V}$ 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

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

and C 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

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

and

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

with

$$\begin{aligned}
\widehat{a_i}^{(k)} &= tr\left(\widehat{u_i\mathbf{y}_i\mathbf{y}_i^\top}^{(k)}\boldsymbol{\Psi}^{-1(k)} - 2\widehat{u_i\mathbf{y}_i\mathbf{z}_i^\top}^{(k)}\boldsymbol{\Lambda}^{\top(k)}\boldsymbol{\Psi}^{-1(k)} + \boldsymbol{\Lambda}^{(k)}\widehat{u_i\mathbf{z}_i\mathbf{z}_i^\top}^{(k)}\boldsymbol{\Lambda}^{\top(k)}\boldsymbol{\Psi}^{-1}\right),\\
\widehat{u_i\mathbf{z}_i\mathbf{z}_i^\top}^{(k)} &= E\left[u_i\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)}\left(\widehat{u_i\mathbf{y}_i\mathbf{y}_i}^{\top(k)} - \widehat{u_i\mathbf{y}_i^{(k)}}\boldsymbol{\beta}^{\top(k)}\mathbf{X}_i^\top\right)\boldsymbol{\Psi}^{-1(k)}\boldsymbol{\Lambda}^{\top(k)}\boldsymbol{\Delta}^{(k)}\\
&+ \boldsymbol{\Delta}^{(k)}\boldsymbol{\Lambda}^{\top(k)}\boldsymbol{\Psi}^{-1(k)}\left(-\mathbf{X}_i\boldsymbol{\beta}^{(k)}\widehat{u_i\mathbf{y}_i}^{(k)} + \mathbf{X}_i\boldsymbol{\beta}^{(k)}\boldsymbol{\beta}^{\top(k)}\mathbf{X}_i^\top\right)\boldsymbol{\Psi}^{-1(k)}\boldsymbol{\Lambda}^{\top(k)}\boldsymbol{\Delta}^{(k)},\\
\boldsymbol{\Delta}^{(k)} &= \left(\boldsymbol{\Omega}^{-1(k)} + \boldsymbol{\Lambda}^{\top(k)}\boldsymbol{\Psi}^{-1(k)}\boldsymbol{\Lambda}^{(k)}\right),\\
\widehat{u_i\mathbf{z}_i}^{(k)} &= E\left[u_i\mathbf{z}_i|\mathbf{V}_i,\mathbf{C}_i,\boldsymbol{\theta}^{(k)}\right] = \boldsymbol{\Delta}^{(k)}\boldsymbol{\Lambda}^{\top(k)}\boldsymbol{\Psi}^{-1(k)}\left(\widehat{u_i\mathbf{y}_i}^{(k)} - \mathbf{X}_i\boldsymbol{\beta}^{(k)}\right),\\
\widehat{u_i\mathbf{y}_i\mathbf{z}_i^\top}^{(k)} &= E\left[u_i\mathbf{y}_i\mathbf{z}_i^\top|\mathbf{V}_i,\mathbf{C}_i,\boldsymbol{\theta}^{(k)}\right] = (\widehat{u_i\mathbf{y}_i\mathbf{y}_i}^{\top(k)} - \widehat{u_i\mathbf{y}_i}^{(k)}\boldsymbol{\beta}^{\top(k)}\mathbf{X}_i^\top)\boldsymbol{\Psi}^{-1(k)}\boldsymbol{\Lambda}^{(k)}\boldsymbol{\Delta}^{(k)}.
\end{aligned}$$

The conditional maximization (CM) steps then conditionally maximize $Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k)})$ with respect to $\boldsymbol{\theta}$ and obtain a new estimate $\widehat{\boldsymbol{\theta}}^{(k+1)}$ as described below:

$$\boldsymbol{\beta}^{(k+1)} = (\sum_{i=1}^{n}\widehat{u}_i^{(k)}\mathbf{X}_i^\top\mathbf{X}_i)^{-1}\sum_{i=1}^{n}\mathbf{X}_i^\top(\widehat{u_i\mathbf{y}_i}^{(k)} - \boldsymbol{\Lambda}^{(k)}\widehat{u_i\mathbf{z}_i}^{(k)}) \tag{8}$$

$$\boldsymbol{\Lambda}_j^{(k+1)} = (\sum_{i=1}^{n}\widehat{u_i\mathbf{z}_i\mathbf{z}_i^\top}^{(k)})^{-1}\sum_{i=1}^{n}\left[(\widehat{u_i\mathbf{y}_i\mathbf{z}_i^\top})_j^{\top(k)} - \widehat{u_i\mathbf{z}_i}^{(k)}(\mathbf{X}_i\boldsymbol{\beta}^{(k)})_j\right] \tag{9}$$

$$\boldsymbol{\Psi}^{(k+1)} = Diag\left(\frac{\sum_{i=1}^{n}(\mathbf{A}_i^{(k)} + \mathbf{A}_i^{(k)\top})}{2n}\right), \tag{10}$$

where $\boldsymbol{\Lambda}_j^\top$ is the j-th row of $\boldsymbol{\Lambda}$ for $j = 1,\ldots,p$ and $\mathbf{A}_i^{(k)} = \widehat{u_i\mathbf{y}_i\mathbf{y}_i^{\top(k)}} - 2\widehat{u_i\mathbf{y}_i}^{(k)}\boldsymbol{\beta}^{\top(k)}\mathbf{X}_i^\top - 2\widehat{u_i\mathbf{y}_i\mathbf{z}_i^\top}^{(k)}\boldsymbol{\Lambda}^{\top(k)} + 2\boldsymbol{\Lambda}^{(k)}\widehat{u_i\mathbf{z}_i}^{(k)}\boldsymbol{\beta}^{\top(k)}\mathbf{X}_i^\top + \mathbf{X}_i\boldsymbol{\beta}^{(k)}\boldsymbol{\beta}^{\top(k)}\mathbf{X}_i^\top + \boldsymbol{\Lambda}^{(k)}\widehat{u\mathbf{z}_i\mathbf{z}_i^\top}^{(k)}\boldsymbol{\Lambda}^{\top(k)}$. The algorithm is iterated until the distance involving two successive evaluations of the 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.

From (8)-(10), it is easy to see that the E-step reduces to the computation of $\widehat{u_i\mathbf{y}_i}^{(k)}$, $\widehat{u_i\mathbf{y}_i\mathbf{y}_i^\top}^{(k)}$ and $\widehat{u_i}^{(k)}$. These expected values can be determined in closed form, using Propositions 1-3, as follows (see Appendix B for details):

- From Proposition 2, if individual i has only censored components:

$$\widehat{u_i\mathbf{y}_i\mathbf{y}_i^\top} = E\left[u_i\mathbf{y}_i\mathbf{y}_i^\top|\mathbf{V}_i,\mathbf{C}_i,\widehat{\boldsymbol{\theta}}\right] = \frac{T_p(\mathbf{V}_i|\widehat{\boldsymbol{\mu}}_i,\widehat{\boldsymbol{\Sigma}}_i^*,\nu+2)}{T_p(\mathbf{V}_i|\widehat{\boldsymbol{\mu}}_i,\widehat{\boldsymbol{\Sigma}}_i^*,\nu)}E\left[\mathbf{W}_i\mathbf{W}_i^\top\right],$$

$$\widehat{u_i\mathbf{y}_i} = E\left[u_i\mathbf{y}_i|\mathbf{V}_i,\mathbf{C}_i,\widehat{\boldsymbol{\theta}}\right] = \frac{T_p(\mathbf{V}_i|\widehat{\boldsymbol{\mu}}_i,\widehat{\boldsymbol{\Sigma}}_i^*,\nu+2)}{T_p(\mathbf{V}_i|\widehat{\boldsymbol{\mu}}_i,\widehat{\boldsymbol{\Sigma}}_i^*,\nu)}E\left[\mathbf{W}_i\right],$$

$$\widehat{u_i} = E\left[u_i|\mathbf{V}_i,\mathbf{C}_i,\widehat{\boldsymbol{\theta}}\right] = \frac{T_p(\mathbf{V}_i|\widehat{\boldsymbol{\mu}}_i,\widehat{\boldsymbol{\Sigma}}_i^*,\nu+2)}{T_p(\mathbf{V}_i|\widehat{\boldsymbol{\mu}}_i,\widehat{\boldsymbol{\Sigma}}_i^*,\nu)},$$

  where $\mathbf{W}_i \sim Tt_p(\widehat{\boldsymbol{\mu}}_i,\widehat{\boldsymbol{\Sigma}}_i^*,\nu+2r,\mathbb{A}_i)$, $\widehat{\boldsymbol{\mu}}_i = \mathbf{X}_i\widehat{\boldsymbol{\beta}}$, $\widehat{\boldsymbol{\Sigma}}_i^* = \dfrac{\nu}{\nu+2}\widehat{\boldsymbol{\Sigma}}_i$, $\widehat{\boldsymbol{\Sigma}}_i = \widehat{\boldsymbol{\Psi}} + \widehat{\boldsymbol{\Lambda}}\widehat{\boldsymbol{\Omega}}\widehat{\boldsymbol{\Lambda}}^\top$ and $\mathbb{A}_i = \{\mathbf{W}_i = (w_1,\ldots,w_p)^\top|w_1 \le V_{i1},\ldots,w_p \le V_{ip}\}$.

- If individual i has only non-censored components, then:

$$\widehat{u\mathbf{y}_i\mathbf{y}_i^\top} = \frac{\nu+p}{\nu+\delta_i}\mathbf{y}_i\mathbf{y}_i^\top,\ \widehat{u\mathbf{y}_i} = \frac{\nu+p}{\nu+\delta_i}\mathbf{y}_i \text{ and } \widehat{u}_i = \frac{\nu+p}{\nu+\delta_i},$$

  where $\delta_i = (\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})^\top\boldsymbol{\Sigma}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})$.

- If individual i has censored and uncensored components, then from Proposition 3 and the fact that $\mathbf{y}_i|\mathbf{V}_i,\mathbf{C}_i$, $\mathbf{y}_i|\mathbf{V}_i,\mathbf{C}_i,\mathbf{y}_i^o$ and $\mathbf{y}_i^c|\mathbf{V}_i,\mathbf{C}_i,\mathbf{y}_i^o$ are equivalent processes, we have:

$$\widehat{u\mathbf{y}_i\mathbf{y}_i^\top} = E\left[u_i\mathbf{y}_i\mathbf{y}_i^\top|\mathbf{y}_i^o,\mathbf{V}_i,\mathbf{C}_i,\hat{\boldsymbol{\theta}}\right] = \begin{pmatrix} \widehat{u}_i\mathbf{y}_i^o\mathbf{y}_i^{o\top} & \widehat{u}_i\mathbf{y}_i^o\widehat{\mathbf{w}_i^c}^\top \\ \widehat{u}_i\widehat{\mathbf{w}_i^c}\mathbf{y}_i^{o\top} & \widehat{u}_i\widehat{\mathbf{w}_i^c\mathbf{w}_i^{c\top}} \end{pmatrix},$$

$$\widehat{u\mathbf{y}}_i = E\left[u_i\mathbf{y}_i|\mathbf{y}_i^o,\mathbf{V}_i,\mathbf{C}_i,\hat{\boldsymbol{\theta}}\right] = vec(\widehat{u}_i\mathbf{y}_i^o,\widehat{\mathbf{w}_i^c}),$$

$$\widehat{u}_i = E\left[u_i|\mathbf{y}_i^o,\mathbf{V}_i,\mathbf{C}_i,\hat{\boldsymbol{\theta}}\right] = \left(\frac{\nu+p_i^o}{\nu+\delta_i^o}\right)\frac{T_{p_i^c}(\mathbf{V}_i|\boldsymbol{\mu}_i^{co},\tilde{\mathbf{S}}_i^{co},\nu+p_i^o+2)}{T_{p_i^c}(\mathbf{V}_i|\boldsymbol{\mu}_i^{co},\boldsymbol{\Sigma}_i^{co},\nu+p_i^o)},$$

  where $\tilde{\mathbf{S}}_i^{co} = \left(\dfrac{\nu+\delta_i^o}{\nu+2+p_i^o}\right)\boldsymbol{\Sigma}_i^{cc.o}$, $\widehat{\mathbf{w}_i^c} = E[\mathbf{W}_i]$ and $\widehat{\mathbf{w}_i^c\mathbf{w}_i^{c\top}} = E[\mathbf{W}_i\mathbf{W}_i^\top]$, with $\mathbf{W}_i \sim Tt_{p_i^c}(\boldsymbol{\mu}_i^{co},\tilde{\mathbf{S}}_i^{co},\nu+p_i^o+2,\mathbb{A}_i^c)$ and $\boldsymbol{\mu}_i^{co}$,$\boldsymbol{\Sigma}_i^{cc.o}$ and $\mathbf{S}_i^{co}$ as defined earlier.

4.1 The variance of the fixed effects

Standard errors of the ML estimates can be approximated by the inverse of the observed information matrix, but there is generally no closed form. Vaida et al. (2007, Sec. 2) suggested using Louis' formula (Louis, 1982) to obtain an adjusted variance-covariance matrix of $\widehat{\boldsymbol{\beta}}$ in the presence of censored information. This method has been also adopted by Vaida and Liu (2009, Sec.2) and Matos et al. (2013b, Sec.3). The estimate of the variance-covariance matrix of $\widehat{\boldsymbol{\beta}}$ is given by the matrix

$$\mathbf{J}_{\boldsymbol{\beta}\boldsymbol{\beta}} = Var(\widehat{\boldsymbol{\beta}}) =$$

$$\left(\sum_{i=1}^{n}\left(\frac{\nu+p}{\nu+p+2}\right)\mathbf{X}_i^{\top}\boldsymbol{\Sigma}_i^{-1}\mathbf{X}_i - \mathbf{X}_i^{\top}\boldsymbol{\Sigma}_i^{-1}Var\left[\left(\frac{\nu+p}{\nu+\delta_i}\right)(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta})|\mathbf{V}_i,\mathbf{C}_i\right]\boldsymbol{\Sigma}_i^{-1}\mathbf{X}_i\right)^{-1},$$

where $\mathbf{y}_i \sim Tt_p(\mathbf{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma}, \nu; \mathbb{A})$ and since $Var\left[\left(\frac{\nu+p}{\nu+\delta_i}\right)(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta})|\mathbf{V}_i,\mathbf{C}_i\right]$ depends on $\widehat{u_i\mathbf{y}_i^2} = E\left[\left(\frac{\nu+p}{\nu+\delta_i}\right)^2\mathbf{y}_i\mathbf{y}_i^{\top}|\mathbf{V}_i,\mathbf{C}_i,\hat{\boldsymbol{\theta}}\right]$, $\widehat{u_i\mathbf{y}_i^1} = E\left[\left(\frac{\nu+p}{\nu+\delta_i}\right)^2\mathbf{y}_i|\mathbf{V}_i,\mathbf{C}_i,\hat{\boldsymbol{\theta}}\right]$ and $\widehat{u_i\mathbf{y}_i^0} = E\left[\left(\frac{\nu+p}{\nu+\delta_i}\right)^2|\mathbf{V}_i,\mathbf{C}_i,\hat{\boldsymbol{\theta}}\right]$, after some algebraic manipulations (see Appendix B), we have

- If individual i has only censored components then from Proposition 2

$$\widehat{u_i\mathbf{y}_i^2} = c_p(\nu,2)\frac{T_p(\mathbf{V}_i|\widehat{\boldsymbol{\mu}}_i,\widehat{\boldsymbol{\Sigma}}_i^{**},\nu+4)}{T_p(\mathbf{V}_i|\widehat{\boldsymbol{\mu}}_i,\widehat{\boldsymbol{\Sigma}}_i^{*},\nu)}E\left[\mathbf{W}_i\mathbf{W}_i^{\top}\right],$$
$$\widehat{u_i\mathbf{y}_i^1} = c_p(\nu,2)\frac{T_p(\mathbf{V}_i|\widehat{\boldsymbol{\mu}}_i,\widehat{\boldsymbol{\Sigma}}_i^{**},\nu+4)}{T_p(\mathbf{V}_i|\widehat{\boldsymbol{\mu}}_i,\widehat{\boldsymbol{\Sigma}}_i^{*},\nu)}E\left[\mathbf{W}_i\right],$$
$$\widehat{u_i\mathbf{y}_i^0} = c_p(\nu,2)\frac{T_p(\mathbf{V}_i|\widehat{\boldsymbol{\mu}}_i,\widehat{\boldsymbol{\Sigma}}_i^{**},\nu+4)}{T_p(\mathbf{V}_i|\widehat{\boldsymbol{\mu}}_i,\widehat{\boldsymbol{\Sigma}}_i^{*},\nu)},$$

where $\mathbf{W}_i \sim Tt_p(\widehat{\boldsymbol{\mu}}_i, \widehat{\boldsymbol{\Sigma}}_i^{*}, \nu+4, \mathbb{A})$, $\widehat{\boldsymbol{\Sigma}}_i^{**} = \frac{\nu}{\nu+4}\widehat{\boldsymbol{\Sigma}}_i$ and

$$c_p(\nu,2) = \left(\frac{\nu+p}{\nu}\right)^2\left[\frac{\Gamma\left(\frac{\nu+p}{2}\right)\Gamma\left(\frac{\nu+4}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\Gamma\left(\frac{\nu+p+4}{2}\right)}\right].$$

- If individual i has only non-censored components, then

$$\widehat{u_i\mathbf{y}_i^2} = \left(\frac{\nu+p}{\nu+\delta_i}\right)^2\mathbf{y}_i\mathbf{y}_i^{\top},\quad \widehat{u_i\mathbf{y}_i^1} = \left(\frac{\nu+p}{\nu+\delta_i}\right)^2\mathbf{y}_i \text{ and } \widehat{u_i\mathbf{y}_i^0} = \left(\frac{\nu+p}{\nu+\delta_i}\right)^2,$$

where $\delta_i = (\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})^{\top}\boldsymbol{\Sigma}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})$.

- If individual i has censored and uncensored components, then from Proposition 3 and the fact that $\mathbf{y}_i|\mathbf{V}_i,\mathbf{C}_i$, $\mathbf{y}_i|\mathbf{V}_i,\mathbf{C}_i,\mathbf{y}_i^o$ and $\mathbf{y}_i^c|\mathbf{V}_i,\mathbf{C}_i,\mathbf{y}_i^o$ are equivalent processes, we have:

$$
\begin{aligned}
\widehat{u_i\mathbf{y}_i^2} &= \begin{pmatrix} \widehat{u_i\mathbf{y}_i^0}\,\mathbf{y}_i^o\mathbf{y}_i^{o\top} & \widehat{u\mathbf{y}_i^0}\,\mathbf{y}_i^o\,\widehat{\mathbf{w}_i^c}^\top \\ \widehat{u_i\mathbf{y}_i^0}\,\widehat{\mathbf{w}_i^c}\,\mathbf{y}_i^{o\top} & \widehat{u\mathbf{y}_i^0}\,\widehat{\mathbf{w}_i^c\,\mathbf{w}_i^{c\top}} \end{pmatrix},\\
\widehat{u\mathbf{y}_i^1} &= vec(\widehat{u_i\mathbf{y}_i^0}\,\mathbf{y}_i^o,\ \widehat{\mathbf{w}_i^c}),\\
\widehat{u\mathbf{y}_i^0} &= \left(\frac{d_p}{(\nu+\delta_i^o)^2}\right)\frac{T_{p_i^c}(\mathbf{V}_i|\boldsymbol{\mu}_i^{co},\tilde{\mathbf{S}}_i^{co},\nu+p_i^o+4)}{T_{p_i^c}(\mathbf{V}_i|\boldsymbol{\mu}_i^{co},\boldsymbol{\Sigma}_i^{co},\nu+p_i^o)},
\end{aligned}
$$

where $d_p=(\nu+p)^2\left(\dfrac{\Gamma((p+\nu)/2)\Gamma((p_1+\nu+4)/2)}{\Gamma((p+\nu)/2)\Gamma((p+\nu+4)/2)}\right)$, $\tilde{\mathbf{S}}_i^{co}=\left(\frac{\nu+\delta_i^o}{\nu+2+p_i^o}\right)\boldsymbol{\Sigma}_i^{cc.o}$, $\widehat{\mathbf{w}_i^c}=E[\mathbf{W}_i]$ and $\widehat{\mathbf{w}_i^c\mathbf{w}_i^{c\top}}=E[\mathbf{W}_i\mathbf{W}_i^\top]$, with $\mathbf{W}_i\sim Tt_{p_i^c}(\boldsymbol{\mu}_i^{co},\tilde{\mathbf{S}}_i^{co},\nu+p_i^o+2,\mathbb{A}_i^c)$ and $\boldsymbol{\mu}_i^{co}$,$\boldsymbol{\Sigma}_i^{cc.o}$ and $\mathbf{S}_i^{co}$ as defined earlier.

Asymptotic confidence intervals and hypothesis tests for the fixed effects are obtained assuming that the ML estimates $\widehat{\boldsymbol{\beta}}$ has approximately an $N_p(\boldsymbol{\beta},\mathbf{J}^{-1}_{\boldsymbol{\beta}\boldsymbol{\beta}})$ distribution. In practice, $\mathbf{J}_{\boldsymbol{\beta}\boldsymbol{\beta}}$ is usually unknown and needs to be replaced by its ML estimates $\mathbf{J}_{\widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{\beta}}}$.

4.2 The estimation of the latent factors

In this subsection, we consider the conditional approach by using the conditional mean to estimate the latent factors (Matos et al., 2013b). Thus, if the values of the parameter vector $\boldsymbol{\theta}=(\boldsymbol{\beta}^\top,\boldsymbol{\Lambda},\boldsymbol{\Psi})^\top$ and ν are known, the conditional mean of $\mathbf{z}_i$ given $\mathbf{V}_i$ and $\mathbf{C}_i$ is:

$$
\begin{aligned}
\widehat{\mathbf{z}}_i(\boldsymbol{\theta}) &= E[\mathbf{z}_i|\mathbf{V}_i,\mathbf{C}_i]=E\left[E[E(\mathbf{z}_i|u_i)|\mathbf{y}_i,u_i]|\mathbf{V}_i,\mathbf{C}_i\right]=E\{\boldsymbol{\Delta}\boldsymbol{\Lambda}^\top\boldsymbol{\Psi}^{-1}\left(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta}\right)|\mathbf{V}_i,\mathbf{C}_i\}\\
&= \boldsymbol{\Delta}\boldsymbol{\Lambda}^\top\boldsymbol{\Psi}^{-1}\left(\widehat{\mathbf{y}}_i-\mathbf{X}_i\boldsymbol{\beta}\right),
\end{aligned}
$$

where $\boldsymbol{\Delta}=\left(\boldsymbol{\Omega}^{-1}+\boldsymbol{\Lambda}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}\right)$ and $\widehat{\mathbf{y}}_i=E\{\mathbf{y}_i|\mathbf{V}_i,\mathbf{C}_i\}$ is the first moment of a truncated multivariate Student-t distribution $Tt_p(\mathbf{X}_i\boldsymbol{\beta},\boldsymbol{\Sigma}_i,\nu;\mathbb{A}_i)$. In practice, the estimators of $\mathbf{z}_i$ can be obtained by substituting the ML estimate $\widehat{\boldsymbol{\theta}}$, which leads to $\widehat{\mathbf{z}_i}=\widehat{\mathbf{z}}_i(\widehat{\boldsymbol{\theta}})$.

The conditional covariance matrix of $\mathbf{z}_i$ given $\mathbf{V}_i$ and $\mathbf{C}_i$ is

$$
\begin{aligned}
&Var[\mathbf{z}_i|\mathbf{V}_i,\mathbf{C}_i]=E[\mathbf{z}_i\mathbf{z}_i^\top|\mathbf{V}_i,\mathbf{C}_i]-\widehat{\mathbf{z}}_i(\boldsymbol{\theta})\widehat{\mathbf{z}}_i(\boldsymbol{\theta})^\top\\
=\ &\boldsymbol{\Delta}E\left[\left(\frac{\nu+p}{\nu+\delta_i}\right)^{-1}|\mathbf{V}_i,\mathbf{C}_i\right]+\boldsymbol{\Delta}\boldsymbol{\Lambda}^\top\boldsymbol{\Psi}^{-1}Var\left[\left(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta}\right)|\mathbf{V}_i,\mathbf{C}_i\right]\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Delta}^\top.
\end{aligned}
$$

These expected values can be easily calculated as a byproduct of our proposed ECM algorithm (E-step).

5 Application

In this section we illustrate the performance of the proposed method first with the analysis of a simulated data set. After that we consider the analysis of the EGRA dataset from Peru.

5.1 Simulation study

The main goal of this simulation study is to evaluate the performance of our proposed method, investigating the effects on the parameter inference when the traditional normality assumption does not hold. In addition, we analyze if the model comparison criteria (AIC and BIC) determine the correct model.

To accomplish our objectives, we present three simulation scenarios under different distributions for the observed data and latent factors as follows:

1. Scenario 1 (TCFA): The data and latent factors have a normal distribution.
2. Scenario 2 (t-TCFA): The data and the latent factors have a Student-t distribution with $\nu = 4$.
3. Scenario 3 (contaminated normal): The data and the latent factors follow a contaminated normal distribution with parameter $\nu = (\nu_1, \nu_2)^\top = (0.1, 0.1)^\top$.

For all scenarios, sample of sizes $n = 50$, 150 and 300 were considered. Under each setting, we fitted the Gaussian TCFA (Subsection 3.1) and the t-TCFA with 4 degrees of freedom (Subsection 3.2) models. Consequently, we have 9 different simulation scenarios with 100 simulated datasets for each one. The MLEs estimates, their associate standard error measures, as well as the AIC and BIC were recorded.

The simulated data are generated as follows:

$$\begin{aligned}
\mathbf{Y}_i|\mathbf{Z}_i, U_i = u_i &\sim N_p(\boldsymbol{\mu}_i, u_i^{-1}\boldsymbol{\Psi}),\\
\mathbf{Z}_i|U_i = u_i &\sim N_q(\mathbf{0}, u_i^{-1}\boldsymbol{\Omega}),\\
U_i &\sim F(\nu),
\end{aligned}$$

where $\boldsymbol{\mu}_i = \mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\Lambda}\mathbf{z}_i$, with $p = 5$, $k = 3$, $q = 2$ and $\boldsymbol{\beta}^\top = (3.5, 2.5, -1.5)$,

$$\boldsymbol{\Lambda}^\top = \begin{pmatrix} -0.6 & -0.6 & -0.6 & 0 & 0 \\ 0 & 0 & 0 & 0.5 & 0.5 \end{pmatrix},$$

$\boldsymbol{\Psi} = diag(0.3, 0.4, 0.6, 0.2, 0.7)$, and $U_i \equiv 1$ for all i for the Scenario 1 (normal case), $U_i \sim \text{Gamma}(2, 2)$ for all i for the Scenario 2 (Student-t case) and $U_i = \begin{cases} \nu_2 & \text{with prob} \quad \nu_1 \\ 1 & \text{with prob} \quad 1-\nu_1 \end{cases}$, for all i, for the Scenario 3 (contaminated normal case).

In order to assure the model identifiability, $\boldsymbol{\Omega}$ was considered fixed at $0.6\mathbf{I}+0.4\mathbf{J}$, where $\mathbf{I}$ is a 2×2 identity matrix and $\mathbf{J}$ is a 2×2 matrix of ones. All generated values of $\mathbf{Y}$ lower than 0 were fixed at this threshold, *i.e.*, the censored values are equal to 0.

The covariate matrices X_i were 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 x_{ij}, $j = 1, 2, 3$. The covariate x_{i1} was equal to 1 for all $i = 1, \ldots, n$ representing the intercept; x_{i2} was generated from a discrete uniform distribution with points 6, 7, 8, 9, 10 and probabilities 0.05, 0.4, 0.45, 0.05, 0.05 respectively, mimicking the variable *age* presented in the real data analysis. Note that for numerical stability, this variable was centered at 8. Finally, the last covariate x_{i3}, was generated from a Bernoulli distribution with paremeter $p = 0.5$, mimicking the variable *gender* in the EGRA data set.

Table 1 presents the summary statistics for parameter estimation assuming that the true model is the Gaussian TCFA for the three considered sample sizes. As is expected, the TCFA model had better performance than the t-TCFA for recovering the true parameter values. Moreover, Table 1 also provides the mean values of the approximate standard errors of the fixed effects estimates obtained through Equation 11 (in parentheses) and the Monte Carlo standard deviation (MC Sd). As can be seen, the theoretically estimated standard errors are relatively close to the Monte Carlo approximations. This result shows that the proposed asymptotic approximation for the variances of the fixed effects is reliable for both TCFA and t-TCFA models. Note that the standard errors decrease as the sample size increases. We also report the Monte Carlo mean of the model comparison criteria mentioned earlier (MC AIC and MC BIC). Note that the MC AIC and MC BIC indicate that the TCFA is the best model.

Table 2 shows the results obtained under the Student-t datasets. Note that the t-TCFA model outperforms the TCFA model as expected. Moreover, the TCFA model

Table 1: Results based on 100 simulated normal datasets. MC mean and MC Sd are the respective Monte Carlo mean and standard deviations of the parameter estimates from fitting the TCFA and the t-TCFA. The numbers in parentheses are the mean values of the approximate standard errors obtained through the information-based method. MC AIC and MC BIC are the Monte Carlo mean of the model comparison measures AIC and BIC respectively.

Sample size	Parameters	TCFA		t-TCFA	
		Mc mean	MC Sd	Mc mean	MC Sd
n=50	β_1	3.495 (0.097)	0.102	3.48 (0.088)	0.104
	β_2	2.507 (0.067)	0.062	2.494 (0.063)	0.062
	β_3	-1.495 (0.135)	0.141	-1.476 (0.124)	0.138
	λ_{11}	-0.546	0.104	-0.512	0.106
	λ_{21}	-0.562	0.127	-0.531	0.128
	λ_{31}	-0.573	0.165	-0.548	0.159
	λ_{42}	0.290	0.147	0.326	0.102
	λ_{52}	0.437	0.216	0.537	0.144
	ψ_{11}	0.317	0.090	0.274	0.082
	ψ_{22}	0.369	0.099	0.319	0.088
	ψ_{33}	0.581	0.167	0.495	0.149
	ψ_{44}	0.335	0.106	0.260	0.077
	ψ_{55}	0.688	0.170	0.493	0.105
	MC AIC	530.745		552.484	
	MC BIC	555.601		577.340	
n=150	β_1	3.493 (0.053)	0.049	3.483 (0.048)	0.050
	β_2	2.500 (0.043)	0.043	2.492 (0.040)	0.046
	β_3	-1.484 (0.079)	0.076	-1.476 (0.070)	0.076
	λ_{11}	-0.563	0.064	-0.531	0.062
	λ_{21}	-0.579	0.081	-0.551	0.080
	λ_{31}	-0.590	0.089	-0.566	0.084
	λ_{42}	0.332	0.088	0.340	0.061
	λ_{52}	0.496	0.118	0.555	0.062
	ψ_{11}	0.312	0.054	0.270	0.050
	ψ_{22}	0.382	0.061	0.328	0.052
	ψ_{33}	0.589	0.089	0.503	0.080
	ψ_{44}	0.312	0.061	0.248	0.044
	ψ_{55}	0.667	0.110	0.494	0.071
	MC AIC	1620.169		1691.247	
	MC BIC	1659.307		1730.386	
n=300	β_1	3.499 (0.039)	0.038	3.488 (0.035)	0.037
	β_2	2.501 (0.032)	0.033	2.491 (0.030)	0.034
	β_3	-1.500 (0.056)	0.050	-1.488 (0.050)	0.049
	λ_{11}	-0.570	0.047	-0.540	0.045
	λ_{21}	-0.580	0.057	-0.554	0.057
	λ_{31}	-0.587	0.062	-0.566	0.060
	λ_{42}	0.339	0.046	0.343	0.034
	λ_{52}	0.511	0.068	0.558	0.044
	ψ_{11}	0.313	0.039	0.274	0.034
	ψ_{22}	0.390	0.051	0.337	0.044
	ψ_{33}	0.589	0.065	0.503	0.059
	ψ_{44}	0.314	0.037	0.254	0.027
	ψ_{55}	0.676	0.068	0.511	0.047
	MC AIC	3228.259		3378.058	
	MC BIC	3276.408		3426.207	

generates larger estimates for the specific variances ($\mathbf{\Psi}$) than the t-TCFA model for all considered sample sizes. Note that as in the normal case, the standard errors are adequately calculated and they decrease as the sample size increases. Finally, the model comparison criteria select the true model (t-TCFA) for all considered sample sizes.

As was mentioned above, in scenario 3 we study the behaviour of the TCFA and t-TCFA models when neither of them are the true generating model. Table 3 indicates that the t-TCFA model fits the contaminated normal datasets better than the TCFA one. The reported values of the MC AIC and MC BIC also confirm this. Figures 1 -5 present the estimates of bias and mean square error (MSE) of each model parameter for the 100 datasets under different sample sizes when the data were generated from scenario 3. Note that, in general, the t-TCFA model presents smaller biases and MSE measures than the TCFA one. This fact assures that the t-TCFA model is more robust to deviations from the model assumptions and fits better than the Gaussian model when neither is the true generating model.

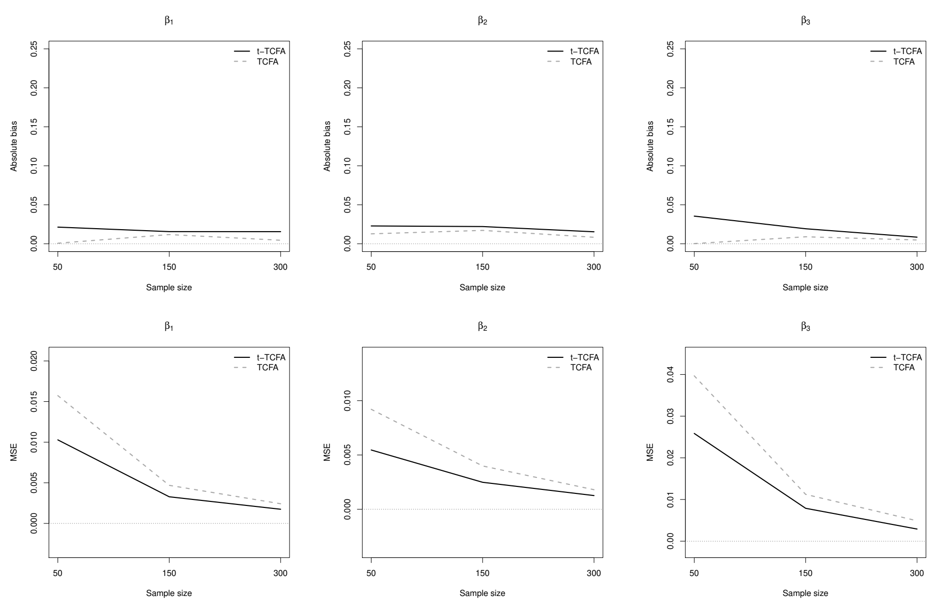

Figure 1: Simulated data. Bias and MSE for the parameter $\boldsymbol{\beta}$ under scenario 3.

Table 2: Results based on 100 simulated Student-t datasets. MC mean and MC Sd are the respective Monte Carlo mean and standard deviations of the parameter estimates from fitting the TCFA and the t-TCFA. The numbers in parentheses are the mean values of the approximate standard errors obtained through the information-based method. MC AIC and MC BIC are the Monte Carlo mean of the model comparison measures AIC and BIC respectively.

Sample size	Parameters	TCFA		t-TCFA	
		Mc mean	MC Sd	Mc mean	MC Sd
n=50	β_1	3.500 (0.133)	0.136	3.481 (0.102)	0.106
	β_2	2.514 (0.091)	0.096	2.480 (0.072)	0.073
	β_3	-1.523 (0.185)	0.190	-1.505 (0.143)	0.149
	λ_{11}	-0.762	0.268	-0.610	0.134
	λ_{21}	-0.786	0.275	-0.621	0.158
	λ_{31}	-0.788	0.332	-0.628	0.196
	λ_{42}	0.416	0.264	0.391	0.143
	λ_{52}	0.607	0.357	0.638	0.170
	ψ_{11}	0.574	0.262	0.358	0.112
	ψ_{22}	0.661	0.300	0.433	0.151
	ψ_{33}	1.037	0.336	0.694	0.209
	ψ_{44}	0.659	0.326	0.339	0.105
	ψ_{55}	1.315	0.685	0.661	0.199
	MC AIC	648.072		639.362	
	MC BIC	672.929		664.218	
n=150	β_1	3.511 (0.073)	0.070	3.485 (0.055)	0.060
	β_2	2.507 (0.059)	0.059	2.484 (0.046)	0.046
	β_3	-1.522 (0.109)	0.100	-1.497 (0.082)	0.082
	λ_{11}	-0.780	0.155	-0.619	0.067
	λ_{21}	-0.821	0.179	-0.647	0.088
	λ_{31}	-0.823	0.225	-0.645	0.115
	λ_{42}	0.484	0.156	0.408	0.078
	λ_{52}	0.731	0.228	0.666	0.088
	ψ_{11}	0.590	0.235	0.349	0.056
	ψ_{22}	0.719	0.219	0.442	0.086
	ψ_{33}	1.024	0.218	0.644	0.105
	ψ_{44}	0.586	0.149	0.324	0.049
	ψ_{55}	1.283	0.319	0.664	0.107
	MC AIC	2004.130		1950.676	
	MC BIC	2043.269		1989.815	
n=300	β_1	3.509 (0.053)	0.054	3.487 (0.040)	0.043
	β_2	2.499 (0.044)	0.046	2.483 (0.034)	0.034
	β_3	-1.507 (0.077)	0.072	-1.492 (0.057)	0.058
	λ_{11}	-0.785	0.110	-0.615	0.053
	λ_{21}	-0.821	0.112	-0.648	0.060
	λ_{31}	-0.824	0.143	-0.652	0.083
	λ_{42}	0.483	0.117	0.404	0.055
	λ_{52}	0.734	0.173	0.657	0.069
	ψ_{11}	0.583	0.143	0.353	0.044
	ψ_{22}	0.724	0.132	0.446	0.059
	ψ_{33}	1.055	0.203	0.649	0.080
	ψ_{44}	0.584	0.115	0.326	0.040
	ψ_{55}	1.283	0.263	0.663	0.077
	MC AIC	3999.002		3885.252	
	MC BIC	4047.151		3933.401	

Table 3: Results based on 100 simulated contaminated normal datasets. MC mean and MC Sd are the respective Monte Carlo mean and standard deviations of the parameter estimates from fitting the TCFA and the t-TCFA. The numbers in parentheses are the mean values of the approximate standard errors obtained through the information-based method. MC AIC and MC BIC are the Monte Carlo mean of the model comparison measures AIC and BIC respectively.

Sample size	Parameters	TCFA		t-TCFA	
		Mc mean	MC Sd	Mc mean	MC Sd
n=50	β_1	3.499 (0.127)	0.126	3.479 (0.098)	0.100
	β_2	2.487 (0.087)	0.096	2.477 (0.070)	0.071
	β_3	-1.500 (0.177)	0.200	-1.465 (0.137)	0.158
	λ_{11}	-0.734	0.266	-0.580	0.128
	λ_{21}	-0.766	0.300	-0.608	0.141
	λ_{31}	-0.742	0.328	-0.597	0.173
	λ_{42}	0.436	0.262	0.383	0.106
	λ_{52}	0.628	0.371	0.601	0.137
	ψ_{11}	0.533	0.244	0.330	0.098
	ψ_{22}	0.624	0.294	0.394	0.108
	ψ_{33}	0.993	0.505	0.594	0.190
	ψ_{44}	0.609	0.274	0.323	0.094
	ψ_{55}	1.255	0.638	0.614	0.186
	MC AIC	638.215		620.283	
	MC BIC	663.071		645.139	
n=150	β_1	3.512 (0.072)	0.068	3.484 (0.053)	0.055
	β_2	2.483 (0.058)	0.061	2.478 (0.044)	0.045
	β_3	-1.509 (0.106)	0.106	-1.481 (0.079)	0.087
	λ_{11}	-0.772	0.150	-0.588	0.076
	λ_{21}	-0.788	0.158	-0.612	0.077
	λ_{31}	-0.773	0.188	-0.628	0.104
	λ_{42}	0.462	0.173	0.391	0.069
	λ_{52}	0.664	0.221	0.617	0.072
	ψ_{11}	0.551	0.170	0.335	0.064
	ψ_{22}	0.667	0.187	0.405	0.077
	ψ_{33}	1.044	0.270	0.609	0.103
	ψ_{44}	0.604	0.152	0.313	0.043
	ψ_{55}	1.212	0.369	0.602	0.095
	MC AIC	1983.951		1903.606	
	MC BIC	2023.089		1942.744	
n=300	β_1	3.505 (0.052)	0.049	3.484 (0.038)	0.039
	β_2	2.492 (0.043)	0.042	2.485 (0.033)	0.032
	β_3	-1.505 (0.075)	0.070	-1.492 (0.055)	0.054
	λ_{11}	-0.762	0.116	-0.590	0.057
	λ_{21}	-0.775	0.112	-0.611	0.058
	λ_{31}	-0.781	0.138	-0.622	0.072
	λ_{42}	0.464	0.119	0.389	0.045
	λ_{52}	0.688	0.165	0.624	0.049
	ψ_{11}	0.568	0.127	0.340	0.042
	ψ_{22}	0.691	0.143	0.410	0.047
	ψ_{33}	1.042	0.204	0.613	0.073
	ψ_{44}	0.577	0.112	0.305	0.033
	ψ_{55}	1.212	0.266	0.610	0.067
	MC AIC	3944.315		3779.822	
	MC BIC	3992.464		3827.971	

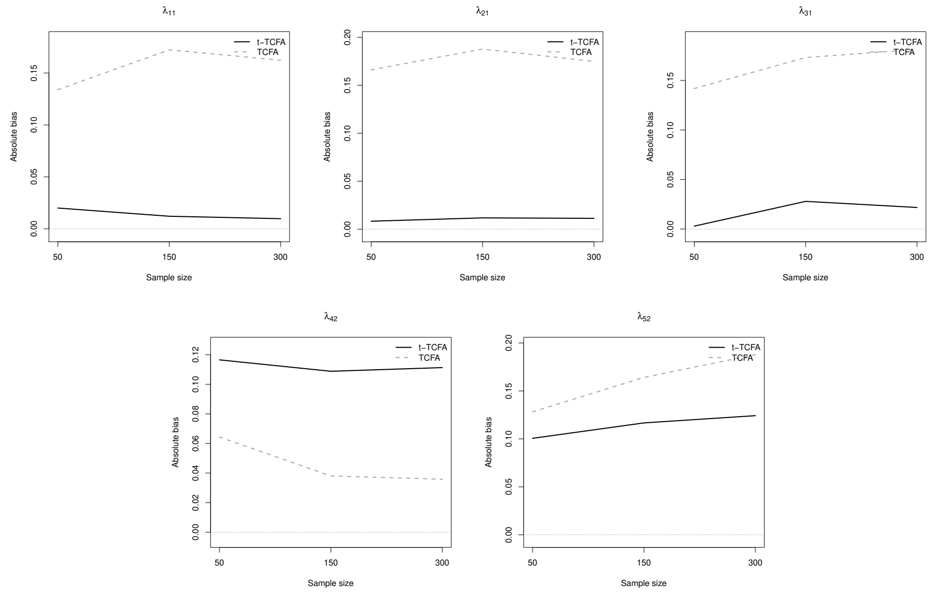

Figure 2: Simulated data. Bias for the parameter $\mathbf{\Lambda}$ under scenario 3.

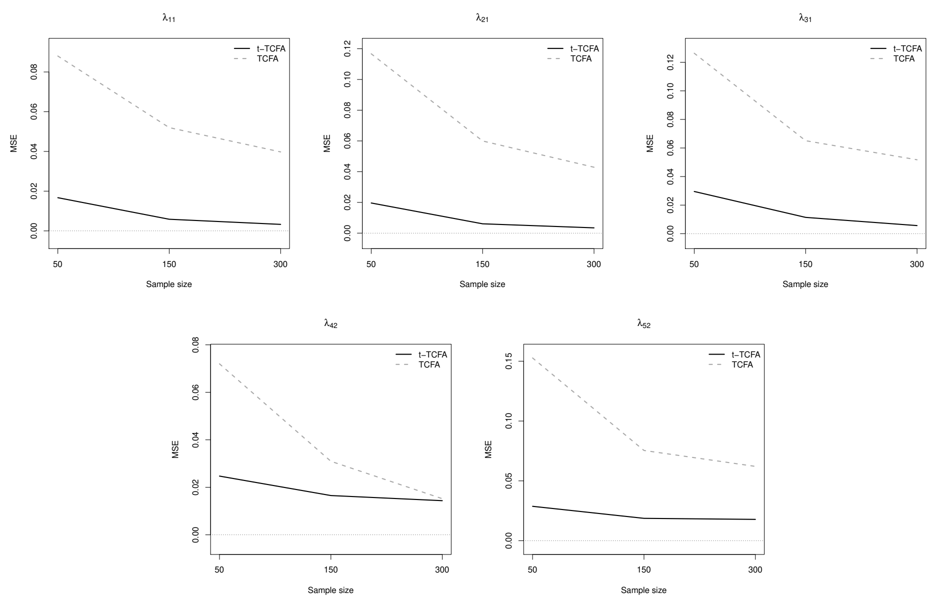

Figure 3: Simulated data. MSE for the parameter $\mathbf{\Lambda}$ under scenario 3.

5.2 Real data

In this section, we analyse the dataset presented in Costa et al. (2013). The data come from a study held in 2007 of Early Grade Reading Assessment (EGRA) test results of 502 Peruvian students. The EGRA test is a simple instrument that reports

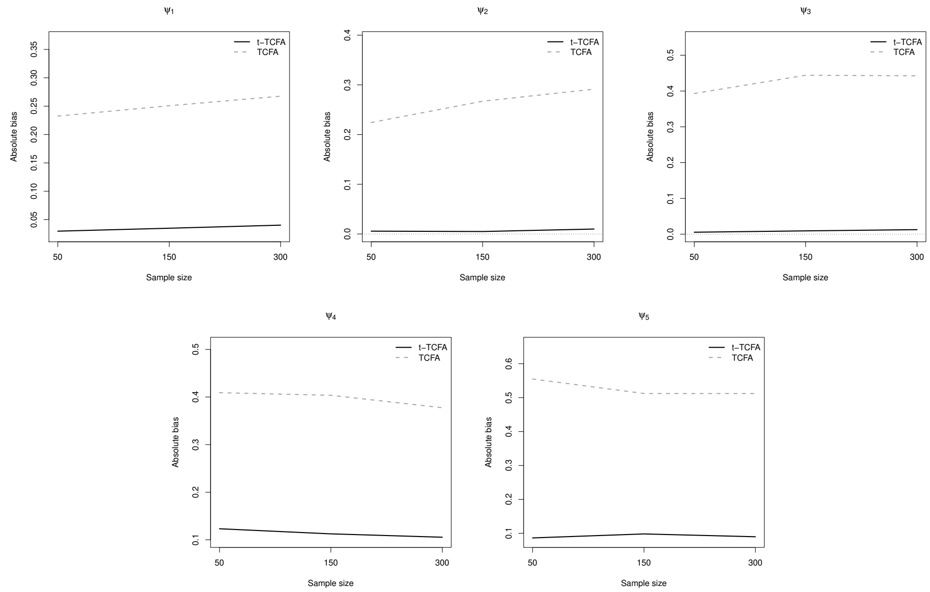

Figure 4: Simulated data. Bias for the parameter $\mathbf{\Psi}$ under scenario 3.

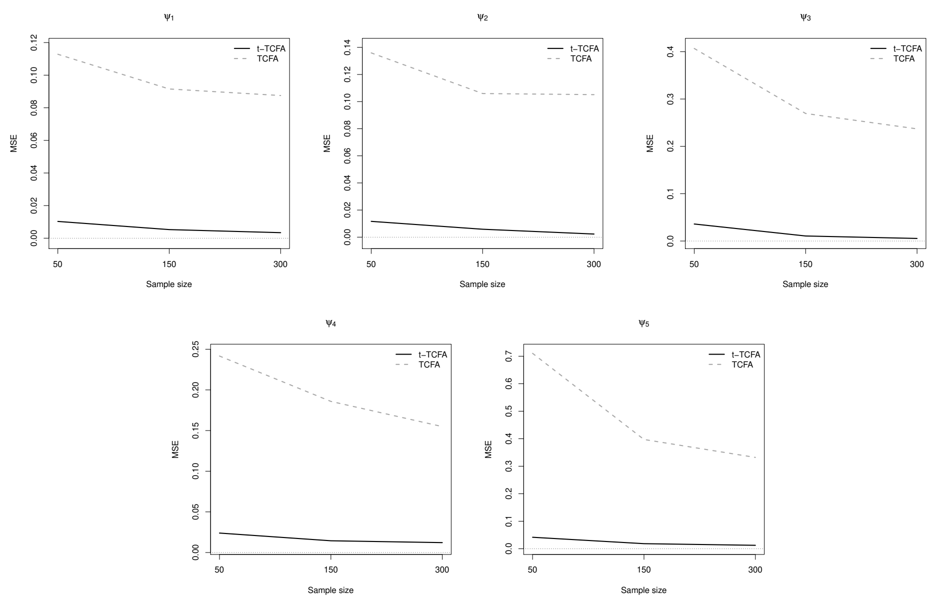

Figure 5: Simulated data. MSE for the parameter $\mathbf{\Psi}$ under scenario 3.

levels of student learning, including assessment of the first steps students take in learning to read, namely, recognizing letters of the alphabet, reading simple words, and understanding sentences and paragraphs. Consequently, this exam evaluates how children in the early grades are acquiring key reading skills. It is an oral test and the time of the responses of each student is recorded. For four out of ten tasks, the required speed to perform these tasks was computed. By taking only one minute in each task, some students present low scores that possibly could have been much better without the time restriction. Like in Costa et al. (2013), we considered the velocity measures of the 10% slowest scores as censored outcomes. Using this definition of censored observations, the TCFA model was defined by

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

where $\mathbf{y}_i$ is a 4×1 vector of the velocity responses given by

y_{i1}**:** ratio between the number of letters of the alphabet recognized by student i and the time spent on this task (under 60 seconds);

y_{i2}**:** ratio between the number of simple words recognized by student i and the time spent on this task (under 60 seconds);

y_{i3}**:** ratio between the number of meaningless words correctly read by student i and the time spent on this task (under 60 seconds);

y_{i4}**:** ratio between the number of correct simple words read in a passage by student i and the time spent on this task (under 60 seconds).

The matrix $\mathbf{X}_i$ corresponds to the 4×5 design matrix of the covariates. The fixed effects vector is $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_4)^\top$, where β_1 represents the gender(0=female, 1=male); β_2 represents the grade (0=second year; 1=third year); β_3 represents the residence zone (0=rural, 1=urban) and β_4 represents the age. $\boldsymbol{\Lambda}$ is a 4×1 vector of factor loadings; $\mathbf{z}_i$ is a latent factor associated with a general ability; $\boldsymbol{\epsilon}_i$ is the vector of random errors of dimension 4×1 and $\boldsymbol{\Omega}$ was set equal to 1.

The total sample size is 502 students and it is divided as follows: 157 are girls and 345 boys; 354 and 148 students come from the second and third grade respectively, and the number of students coming from urban and rural zone are 250 and 252 respectively. Also, 51% of students from the sample were seven years old or less. The performance of these students in each task is presented in Figure 6. As can be seen, these variables have from moderate to high correlations. Since the sample presents censored outcomes, the use of a tobit CFA model is well justified.

Similarly to the empirical studies, we conducted an analysis comparing the Gaussian TCFA and t-TCFA models. In Figure 7, we plot the log-likelihood of the t-TCFA model for different values of ν. Note that the maximum of the likelihood was

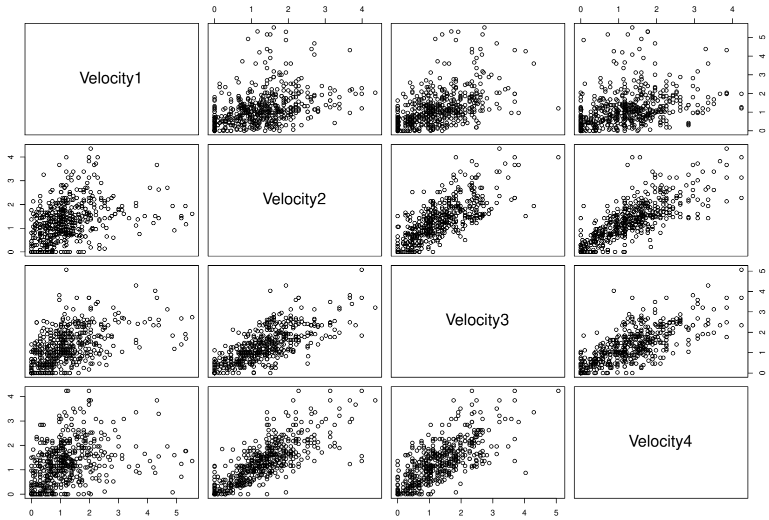

Figure 6: Scatterplots of EGRA velocity measures.

at $\nu = 5$. Consequently and as recommended by Lange et al. (1989), we fixed the parameter ν at this value. In Table 4, we summarize the MLEs for $\boldsymbol{\beta}$ with their standard errors in parentheses, $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$ for both models. The response variables were standardized to avoid differences in the measurement scales.

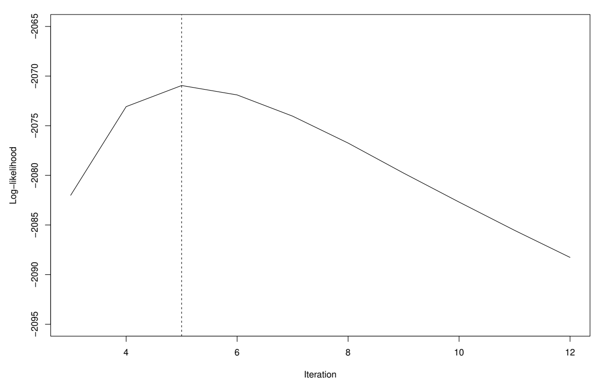

Figure 7: EGRA dataset. Profile likelihood for different values of ν

Table 4 shows that, in general, the estimates corresponding to the t-TCFA model are smaller than those obtained under the Gaussian TCFA. Both models indicate that all variables (gender, grade, zone residence and age) are significantly different from zero and aside from the small value of factor loading in the velocity scores in Task 1 (“recognizing letters of the alphabet”), the factor loadings confirm our

Table 4: EGRA dataset. ML estimates under TCFA and t-TCFA models. Standard errors for the fixed effects are in parentheses.

Parameters	t-TCFA	TCFA
β_1	0.276 (0.068)	0.251 (0.093)
β_2	0.655 (0.086)	0.765 (0.119)
β_3	0.785 (0.057)	0.931 (0.079)
β_4	-0.111 (0.044)	-0.123 (0.06)
λ_{11}	0.479	0.565
λ_{21}	0.880	0.967
λ_{31}	0.815	0.915
λ_{41}	0.885	0.961
ψ_{11}	0.457	0.788
ψ_{22}	0.114	0.161
ψ_{33}	0.172	0.224
ψ_{44}	0.138	0.206
AIC	4165.902	4231.904
BIC	4216.525	4282.527
CAIc	4228.525	4294.527

initial hypothesis of a general latent factor. This factor was interpreted as "fluency in Spanish" and, as expected, students with high velocity measures have high scores in this factor. It is also interesting to point out that the specific variance associated with Task 1 was smaller for the t-TCFA model. This fact can indicate better fit of this measure to this model compared to the classic TCFA. The model selection criteria also indicate that the t-TCFA model outperforms the Gaussian one.

In order to assess the convergence of the proposed ECM algorithm, we compare the performance of the t-TCFA and TCFA loglikelihood values with respect to the iterations of that algorithm. The results shown in Figure 8. As expected, the likelihood values stabilize as the number of iteration increases and both models reach convergence. Note that the t-TCFA model has better performance than the Gaussian one for a higher number of iterations.

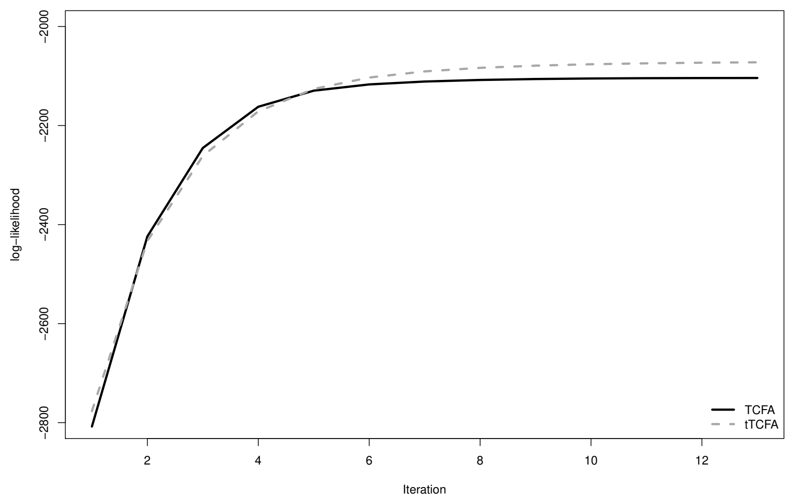

Figure 8: EGRA dataset. Convergence of the log-likelihoods between the two ECM algorithms for each iteration.

6 Conclusions

In this paper we proposed an exact EM-type algorithm for estimating a multivariate tobit confirmatory factor analysis with covariates based on the Student-t distribution. The proposed robust model is particularly useful in analysis of multivariate censored data with presence of outliers. The exact EM-type algorithm uses closed form expressions at the E-step, as opposed to the Monte Carlo EM algorithm proposed by Zhou and Liu (2009). These expressions rely on formulas of the mean and variance of a truncated multivariate Student-t distribution. Analytical expressions for these moments were derived by Ho et al. (2012) and they require the computation of the multivariate Student-t *cdf*. This task can be done efficiently by using the *mvtnorm* package available in `R`. The simulation study and the analysis of the

EGRA data set showed that the t-TCFA outperforms the traditional Gaussian tobit CFA model under the presence of data coming from a heavy-tailed distribution. The likelihood function is derived with no additional computational cost, allowing practitioners to implement different model selection procedures. Since the EM likelihood sequence is monotonic, the difficulties in assessing convergence, typically present in many MCMC algorithms, are avoided.

We believe this paper introduces a novel method for the analysis of CFA models that can be easily implemented by practitioners in other areas where censored information appears frequently, for example, multivariate measurement error models, spatial models, space state models, among many others. In this context, the proposed EM algorithm has been coded and implemented completely in `R` and the script is available from us upon request.

Finally, our model can be extended by assuming that the latent factors and unobservable errors follow a jointly skewed distribution, as was considered by Lin et al. (2013). This extension is currently under investigation and will be part of a future communication.

Appendix A: Proofs of Propositions

Proof of Proposition 1: The proof of (i) is straightforward from Equation (1). The proof of (ii) follows from Proposition 4 given in Arellano-Valle and Genton (2010) by setting $\lambda = \tau = 0$.

Proof of Proposition 2: First note that if $\mathbf{X} \sim t_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$, then we can write

$$\left(\frac{\nu+p}{\nu+\delta}\right)^r t_p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu) = c_p(\nu, r) t_p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}^*, \nu+2r).$$

It follows that

$$E\left\{\left(\frac{\nu+p}{\nu+\delta}\right)^r \mathbf{X}^{(k)}\right\} = c_p(\nu, r)\frac{T_p(\mathbf{a}|\boldsymbol{\mu}, \boldsymbol{\Sigma}^*, \nu+2r)}{T_p(\mathbf{a}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)} E\left\{\mathbf{X}^{(k)}|\mathbf{X} \leq \mathbf{a}\right\},$$

$\int_{\mathbf{w}\leq\mathbf{a}} \mathbf{w}^{(k)} \frac{t_p(\mathbf{w}|\boldsymbol{\mu}, \boldsymbol{\Sigma}^*, \nu+2)}{T_p(\mathbf{a}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)} d\mathbf{w} = \frac{T_p(\mathbf{a}|\boldsymbol{\mu}, \boldsymbol{\Sigma}^*, \nu+2)}{T_p(\mathbf{a}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)} \int_{\mathbf{w}\leq\mathbf{a}} \frac{t_p(\mathbf{w}|\boldsymbol{\mu}, \boldsymbol{\Sigma}^*, \nu+2)}{T_p(\mathbf{a}|\boldsymbol{\mu}, \boldsymbol{\Sigma}^*, \nu+2)} d\mathbf{w}$, which concludes the proof.

Proof of Proposition 3: First note that if $\mathbf{X} \sim t_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$. Then using the result given in Proposition 1-(ii), we have

$$\left(\frac{\nu+p}{\nu+\delta}\right)^r t_{p_2}\left(\mathbf{x}_2|\boldsymbol{\mu}_{2.1}, \widetilde{\boldsymbol{\Sigma}}_{22.1}, \nu+p_1\right) = \frac{d_p(p_1, \nu, r)}{(\nu+\delta_1)^r} t_{p_2}(\mathbf{x_2}|\boldsymbol{\mu}_{2.1}, \widetilde{\boldsymbol{\Sigma}}^*_{22.1}, \nu+p_1+2r)$$

and the proof concludes by noting that

$$E\left\{\left(\frac{\nu+p}{\nu+\delta}\right)^r \mathbf{X}_2^{(k)}|\mathbf{X}_1\right\} = \frac{d_p(p_1,\nu,r)}{(\nu+\delta_1)^r}\frac{T_{p_2}(\mathbf{a}^{x_2}|\boldsymbol{\mu}_{2.1},\widetilde{\boldsymbol{\Sigma}}^*_{22.1},\nu+p_1+2r)}{T_{p_2}(\mathbf{a}^{x_2}|\boldsymbol{\mu}_{2.1},\widetilde{\boldsymbol{\Sigma}}_{22.1},\nu+p_1)}E\left\{\mathbf{X}_2^{(k)}|\mathbf{X}_2 \leq \mathbf{a}^{x_2}\right\},$$

where $\mathbf{X}_2 \sim t_{p_2}\left(\boldsymbol{\mu}_{2.1},\widetilde{\boldsymbol{\Sigma}}^*_{22.1},\nu+p_1+2r\right)$.

Appendix B: Details of the EM algorithm

First, we establish the following Lemma, which will be useful in our procedures. Its proof can be found in Arellano-Valle et al. (2005).

Lemma 1. *Let* $\mathbf{Y} \overset{\text{ind.}}{\sim} N_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$ *and* $\mathbf{x} \overset{\text{ind.}}{\sim} N_q(\boldsymbol{\eta},\boldsymbol{\Omega})$. *So,*

$$\begin{aligned}\phi_p(\mathbf{y}|\boldsymbol{\mu}+x,\boldsymbol{\Sigma})\phi_q(x,\boldsymbol{\Omega}) &= \phi_p(\mathbf{y}|\boldsymbol{\mu}+\boldsymbol{\eta},\boldsymbol{\Sigma}+\boldsymbol{\Omega}^\top)\\ &\quad\times\phi_q(x|\boldsymbol{\eta}+\boldsymbol{\Delta}^\top\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu}-\boldsymbol{\eta}),\boldsymbol{\Sigma}),\end{aligned}$$

where $\boldsymbol{\Delta} = (\boldsymbol{\Omega}^{-1}+\boldsymbol{\Sigma}^{-1})^{-1}$.

Let $\mathbf{y} = (\mathbf{y}_1^\top,\ldots,\mathbf{y}_n^\top)^\top$, $\mathbf{z} = (\mathbf{z}_1^\top,\ldots,\mathbf{z}_n^\top)^\top$, $\mathbf{u} = (u_1,\ldots,u_n)^\top$, $\mathbf{V} = vec(\mathbf{V}_1,\ldots,\mathbf{V}_n)$ and $\mathbf{C} = vec(\mathbf{C}_1,\ldots,\mathbf{C}_n)$, such that we observe $(\mathbf{V}_i,\mathbf{C}_i)$ for the ith subject. In the estimation procedure, $\mathbf{z}$, $\mathbf{u}$ and $\mathbf{y}$ are treated as hypothetical missing data and augmented with the observed data set $\mathbf{y}_c = (\mathbf{C}^\top,\mathbf{V}^\top,\mathbf{y}^\top,\mathbf{z}^\top,\mathbf{u}^\top)^\top$. Thus, we have

$$L(\mathbf{y}_c|\boldsymbol{\theta}) = \prod_{i=1}^n f(\mathbf{y}_i,\mathbf{z}_i,u_i) = \prod_{i=1}^n f(\mathbf{y}_i|\mathbf{V}_i,\mathbf{C}_i,\mathbf{z}_i,u_i)f(\mathbf{z}_i|u_i)f(u_i).$$

The complete data log-likelihood is given by

$$\begin{aligned}\ell_c(\boldsymbol{\theta}|\mathbf{y}_c) = C &- \frac{1}{2}\sum_{i=1}^n\left[\log|\boldsymbol{\Psi}| + u_i(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta}-\boldsymbol{\Lambda}\mathbf{z}_i)^\top\boldsymbol{\Psi}^{-1}(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta}-\boldsymbol{\Lambda}\mathbf{z}_i)+\log|\boldsymbol{\Omega}|\right]\\ &- \frac{1}{2}\sum_{i=1}^n\left[u_i\mathbf{z}_i^\top\boldsymbol{\Omega}^{-1}\mathbf{z}_i + h(u_i|\nu)\right],\end{aligned}$$

The $Q(\cdot)$ function is given by

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}) = E[\ell_c(\boldsymbol{\theta}|\mathbf{y}_c)|\mathbf{V},\mathbf{C},\boldsymbol{\theta}^{(k)}].$$

So we have

$$\begin{aligned}Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}) &= C^{(k)} - \frac{1}{2}\sum_{i=1}^n\Big\{\log|\boldsymbol{\Psi}| + \log|\boldsymbol{\Omega}| + \mathrm{tr}\left(E[u_i\mathbf{z}_i\mathbf{z}_i^\top|\mathbf{V}_i,\mathbf{C}_i,\boldsymbol{\theta}^{(k)}]\boldsymbol{\Omega}^{-1}\right)\\ &\quad+\mathrm{tr}\left(E\left[u_i\boldsymbol{\Psi}^{-1}(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta}-\boldsymbol{\Lambda}\mathbf{z}_i)(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta}-\boldsymbol{\Lambda}\mathbf{z}_i)^\top|\mathbf{V}_i,\mathbf{C}_i,\boldsymbol{\theta}^{(k)}\right]\right)\Big\},\end{aligned}$$

where $C^{(k)}$ is a constant that is independent of the parameter vector $\boldsymbol{\theta}$.

Thus, to compute the expectation term above, note first that,

$$\begin{aligned} \mathbf{y}_i &\overset{\text{ind.}}{\sim} Tt_p(\mathbf{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma}_i, \nu), \\ E(u_i|\mathbf{y}_i) &= \frac{\nu + p}{\nu + \delta_i}, \end{aligned}$$

where $\delta_i = (\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})^\top \boldsymbol{\Sigma}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})$, and using the Lemma 1, we have

$$\mathbf{z}_i|\mathbf{y}_i, u_i \overset{\text{ind.}}{\sim} N_q\left(\boldsymbol{\Delta}\boldsymbol{\Lambda}^\top\boldsymbol{\Psi}^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}), \boldsymbol{\Delta}u_i^{-1}\right),$$

with $\boldsymbol{\Delta} = (\boldsymbol{\Omega}^{-1} + \boldsymbol{\Lambda}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda})^{-1}$. Using Propositions 1-3, we compute the following expectation terms:

$$\begin{aligned} \widehat{u_i\mathbf{y}_i} &= E\{u_i\mathbf{y}_i|\mathbf{V}_i, \mathbf{C}_i, \widehat{\boldsymbol{\theta}}\} = E\{E[E(u_i\mathbf{y}_i|\mathbf{y}_i, u_i, \mathbf{V}_i, \mathbf{C}_i, \widehat{\boldsymbol{\theta}})|\mathbf{y}_i, \mathbf{V}_i, \mathbf{C}_i, \widehat{\boldsymbol{\theta}}]|\mathbf{V}_i, \mathbf{C}_i, \widehat{\boldsymbol{\theta}}\} \\ &= E\left[E(u_i\mathbf{y}_i|\mathbf{y}_i, \mathbf{V}_i, \mathbf{C}_i, \widehat{\boldsymbol{\theta}})|\mathbf{V}_i, \mathbf{C}_i, \widehat{\boldsymbol{\theta}}\right] = E\left(\frac{(\nu+p)}{(\nu+\delta_i)}\mathbf{y}_i\middle|\mathbf{V}_i, \mathbf{C}_i, \widehat{\boldsymbol{\theta}}\right) \\ &= \frac{T_p(\mathbf{a}|\boldsymbol{\mu}, \boldsymbol{\Sigma}^*, \nu+2)}{T_p(\mathbf{a}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)}E\{\mathbf{W}_i\}. \end{aligned}$$

$$\begin{aligned} \widehat{u_i\mathbf{y}_i\mathbf{y}_i^\top} &= E\{u_i\mathbf{y}_i\mathbf{y}_i^\top|\mathbf{V}_i, \mathbf{C}_i, \widehat{\boldsymbol{\theta}}\} \\ &= E\{E[E(u_i\mathbf{y}_i\mathbf{y}_i^\top|\mathbf{y}_i, u_i, \mathbf{V}_i, \mathbf{C}_i, \widehat{\boldsymbol{\theta}})|\mathbf{y}_i, \mathbf{V}_i, \mathbf{C}_i, \widehat{\boldsymbol{\theta}}]|\mathbf{V}_i, \mathbf{C}_i, \widehat{\boldsymbol{\theta}}\} \\ &= E\left[E(u_i\mathbf{y}_i\mathbf{y}_i^\top|\mathbf{y}_i, \mathbf{V}_i, \mathbf{C}_i, \widehat{\boldsymbol{\theta}})|\mathbf{V}_i, \mathbf{C}_i, \widehat{\boldsymbol{\theta}}\right] = E\left(\frac{(\nu+p)}{(\nu+\delta_i)}\mathbf{y}_i\mathbf{y}_i^\top\middle|\mathbf{V}_i, \mathbf{C}_i, \widehat{\boldsymbol{\theta}}\right) \\ &= \frac{T_p(\mathbf{a}|\boldsymbol{\mu}, \boldsymbol{\Sigma}^*, \nu+2)}{T_p(\mathbf{a}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)}E\{\mathbf{W}_i\mathbf{W}_i^\top\}. \end{aligned}$$

$$\begin{aligned} \widehat{u}_i &= E\{u_i|\mathbf{V}_i, \mathbf{C}_i, \widehat{\boldsymbol{\theta}}\} = E\{E[E(u_i|\mathbf{y}_i, u_i, \mathbf{V}_i, \mathbf{C}_i, \widehat{\boldsymbol{\theta}})|\mathbf{y}_i, \mathbf{V}_i, \mathbf{C}_i, \widehat{\boldsymbol{\theta}}]|\mathbf{V}_i, \mathbf{C}_i, \widehat{\boldsymbol{\theta}}\} \\ &= E\left[E(u_i|\mathbf{y}_i, \mathbf{V}_i, \mathbf{C}_i, \widehat{\boldsymbol{\theta}})|\mathbf{V}_i, \mathbf{C}_i, \widehat{\boldsymbol{\theta}}\right] = E\left(\frac{(\nu+p)}{(\nu+\delta_i)}\middle|\mathbf{V}_i, \mathbf{C}_i, \widehat{\boldsymbol{\theta}}\right) \\ &= \frac{T_p(\mathbf{a}|\boldsymbol{\mu}, \boldsymbol{\Sigma}^*, \nu+2)}{T_p(\mathbf{a}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)}E\{\mathbf{W}_i^0\} = \frac{T_p(\mathbf{a}|\boldsymbol{\mu}, \boldsymbol{\Sigma}^*, \nu+2)}{T_p(\mathbf{a}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)}. \end{aligned}$$

$$
\begin{aligned}
\widehat{u_i\mathbf{z}_i} &= E\{u_i\mathbf{z}_i|\mathbf{V}_i,\mathbf{C}_i,\boldsymbol{\theta}^{(k)}\}\\
&= E\{E[E(u_i\mathbf{z}_i|\mathbf{y}_i,u_i,\mathbf{V}_i,\mathbf{C}_i,\boldsymbol{\theta}^{(k)})|\mathbf{y}_i,\mathbf{V}_i,\mathbf{C}_i,\boldsymbol{\theta}^{(k)}]|\mathbf{V}_i,\mathbf{C}_i,\boldsymbol{\theta}^{(k)}\}\\
&= E\{E[u_iE(\mathbf{z}_i|\mathbf{y}_i,u_i,\mathbf{V}_i,\mathbf{C}_i,\boldsymbol{\theta}^{(k)})|\mathbf{y}_i,\mathbf{V}_i,\mathbf{C}_i,\boldsymbol{\theta}^{(k)}]|\mathbf{V}_i,\mathbf{C}_i,\boldsymbol{\theta}^{(k)}\}\\
&= E\{E\left[u_i\boldsymbol{\Delta}\boldsymbol{\Lambda}^\top\boldsymbol{\Psi}^{-1}(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta})|\mathbf{y}_i,\mathbf{V}_i,\mathbf{C}_i,\boldsymbol{\theta}^{(k)}\right]|\mathbf{V}_i,\mathbf{C}_i,\boldsymbol{\theta}^{(k)}\}\\
&= E\left[\boldsymbol{\Delta}\boldsymbol{\Lambda}^\top\boldsymbol{\Psi}^{-1}(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta})E(u_i|\mathbf{y}_i,\mathbf{V}_i,\mathbf{C}_i,\boldsymbol{\theta}^{(k)})|\mathbf{V}_i,\mathbf{C}_i,\boldsymbol{\theta}^{(k)}\right]\\
&= E\left\{\boldsymbol{\Delta}\boldsymbol{\Lambda}^\top\boldsymbol{\Psi}^{-1}\left[\left(\frac{(\nu+p)}{(\nu+\delta)}\mathbf{y}_i\right)-\mathbf{X}_i\boldsymbol{\beta}\left(\frac{(\nu+p)}{(\nu+\delta)}\right)\right]\Big|\mathbf{V}_i,\mathbf{C}_i,\boldsymbol{\theta}^{(k)}\right\}\\
&= \boldsymbol{\Delta}\boldsymbol{\Lambda}^\top\boldsymbol{\Psi}^{-1}\left[E\left(\frac{(\nu+p)}{(\nu+\delta)}\mathbf{y}_i\Big|\mathbf{V}_i,\mathbf{C}_i,\boldsymbol{\theta}^{(k)}\right)-\mathbf{X}_i\boldsymbol{\beta}E\left(\frac{(\nu+p)}{(\nu+\delta)}\Big|\mathbf{V}_i,\mathbf{C}_i,\boldsymbol{\theta}^{(k)}\right)\right]\\
&= \boldsymbol{\Delta}\boldsymbol{\Lambda}^\top\boldsymbol{\Psi}^{-1}\left[\frac{T_p(\mathbf{a}|\boldsymbol{\mu},\boldsymbol{\Sigma}^*,\nu+2)}{T_p(\mathbf{a}|\boldsymbol{\mu},\boldsymbol{\Sigma},\nu)}E\{\mathbf{W}_i\}-\mathbf{X}_i\boldsymbol{\beta}\frac{T_p(\mathbf{a}|\boldsymbol{\mu},\boldsymbol{\Sigma}^*,\nu+2)}{T_p(\mathbf{a}|\boldsymbol{\mu},\boldsymbol{\Sigma},\nu)}\right]\\
&= \boldsymbol{\Delta}\boldsymbol{\Lambda}^\top\boldsymbol{\Psi}^{-1}\left[\widehat{u_i\mathbf{y}_i}-\mathbf{X}_i\boldsymbol{\beta}\widehat{u}_i\right].
\end{aligned}
$$

$$
\begin{aligned}
\widehat{u_i\mathbf{z}_i\mathbf{z}_i^\top} &= E\{u_i\mathbf{z}_i\mathbf{z}_i^\top|\mathbf{V}_i,\mathbf{C}_i,\boldsymbol{\theta}^{(k)}\}\\
&= E\{E[E(u_i\mathbf{z}_i\mathbf{z}_i^\top|\mathbf{y}_i,u_i,\mathbf{V}_i,\mathbf{C}_i,\boldsymbol{\theta}^{(k)})|\mathbf{y}_i,\mathbf{V}_i,\mathbf{C}_i,\boldsymbol{\theta}^{(k)}]|\mathbf{V}_i,\mathbf{C}_i,\boldsymbol{\theta}^{(k)}\}\\
&= E\{E[u_iE(\mathbf{z}_i\mathbf{z}_i^\top|\mathbf{y}_i,u_i,\mathbf{V}_i,\mathbf{C}_i,\boldsymbol{\theta}^{(k)})|\mathbf{y}_i,\mathbf{V}_i,\mathbf{C}_i,\boldsymbol{\theta}^{(k)}]|\mathbf{V}_i,\mathbf{C}_i,\boldsymbol{\theta}^{(k)}\}\\
&= E\left\{E\left[u_i\left(u_i^{-1}\boldsymbol{\Delta}+\boldsymbol{\Delta}\boldsymbol{\Lambda}^\top\boldsymbol{\Psi}^{-1}(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta})(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta})^\top\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Delta}^\top\right)\big|\mathbf{y}_i,\mathbf{V}_i,\mathbf{C}_i,\boldsymbol{\theta}^{(k)}\right]\big|\mathbf{V}_i,\mathbf{C}_i,\boldsymbol{\theta}^{(k)}\right.\\
&= \boldsymbol{\Delta}+\boldsymbol{\Delta}\boldsymbol{\Lambda}^\top\boldsymbol{\Psi}^{-1}E\left[E(u_i|\mathbf{y}_i,\mathbf{V}_i,\mathbf{C}_i,\boldsymbol{\theta}^{(k)})(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta})(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta})^\top\big|\mathbf{V}_i,\mathbf{C}_i,\boldsymbol{\theta}^{(k)}\right]\\
&= \boldsymbol{\Delta}+\boldsymbol{\Delta}\boldsymbol{\Lambda}^\top\boldsymbol{\Psi}^{-1}\left\{E\left(\frac{(\nu+p)}{(\nu+\delta)}\mathbf{y}_i\mathbf{y}_i^\top\Big|\mathbf{V}_i,\mathbf{C}_i,\boldsymbol{\theta}^{(k)}\right)-E\left(\frac{(\nu+p)}{(\nu+\delta)}\mathbf{y}_i\Big|\mathbf{V}_i,\mathbf{C}_i,\boldsymbol{\theta}^{(k)}\right)\boldsymbol{\beta}^\top\mathbf{X}_i^\top\right.\\
&\quad\left.-\mathbf{X}_i\boldsymbol{\beta}\left[E\left(\frac{(\nu+p)}{(\nu+\delta)}\mathbf{y}_i\Big|\mathbf{V}_i,\mathbf{C}_i,\boldsymbol{\theta}^{(k)}\right)\right]^\top+E\left(\frac{(\nu+p)}{(\nu+\delta)}\Big|\mathbf{V}_i,\mathbf{C}_i,\boldsymbol{\theta}^{(k)}\right)\mathbf{X}_i\boldsymbol{\beta}\boldsymbol{\beta}^\top\mathbf{X}_i^\top\right\}\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Delta}^\top\\
&= \boldsymbol{\Delta}+\boldsymbol{\Delta}\boldsymbol{\Lambda}^\top\boldsymbol{\Psi}^{-1}\left\{\frac{T_p(\mathbf{a}|\boldsymbol{\mu},\boldsymbol{\Sigma}^*,\nu+2)}{T_p(\mathbf{a}|\boldsymbol{\mu},\boldsymbol{\Sigma},\nu)}E\{\mathbf{W}_i\mathbf{W}_i^\top\}-\frac{T_p(\mathbf{a}|\boldsymbol{\mu},\boldsymbol{\Sigma}^*,\nu+2)}{T_p(\mathbf{a}|\boldsymbol{\mu},\boldsymbol{\Sigma},\nu)}E\{\mathbf{W}_i\}\boldsymbol{\beta}^\top\mathbf{X}_i^\top\right.\\
&\quad\left.-\mathbf{X}_i\boldsymbol{\beta}\left[\frac{T_p(\mathbf{a}|\boldsymbol{\mu},\boldsymbol{\Sigma}^*,\nu+2)}{T_p(\mathbf{a}|\boldsymbol{\mu},\boldsymbol{\Sigma},\nu)}E\{\mathbf{W}_i\}\right]^\top+\frac{T_p(\mathbf{a}|\boldsymbol{\mu},\boldsymbol{\Sigma}^*,\nu+2)}{T_p(\mathbf{a}|\boldsymbol{\mu},\boldsymbol{\Sigma},\nu)}\mathbf{X}_i\boldsymbol{\beta}\boldsymbol{\beta}^\top\mathbf{X}_i^\top\right\}\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Delta}^\top\\
&= \boldsymbol{\Delta}+\boldsymbol{\Delta}\boldsymbol{\Lambda}^\top\boldsymbol{\Psi}^{-1}\left(\widehat{u_i\mathbf{y}_i\mathbf{y}_i^\top}-\widehat{u_i\mathbf{y}_i}\boldsymbol{\beta}^\top\mathbf{X}_i^\top-\mathbf{X}_i\boldsymbol{\beta}\widehat{u_i\mathbf{y}_i}^\top+\widehat{u}_i\mathbf{X}_i\boldsymbol{\beta}\boldsymbol{\beta}^\top\mathbf{X}_i^\top\right)\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Delta}^\top.
\end{aligned}
$$

$$
\begin{aligned}
\widehat{u_i\mathbf{y}_i\mathbf{z}_i} &= E\{u_i\mathbf{y}_i\mathbf{z}_i^\top|\mathbf{V}_i,\mathbf{C}_i,\boldsymbol{\theta}^{(k)}\}\\
&= E\{E[E(u_i\mathbf{y}_i\mathbf{z}_i^\top|\mathbf{y}_i,u_i,\mathbf{V}_i,\mathbf{C}_i,\boldsymbol{\theta}^{(k)})|\mathbf{y}_i,\mathbf{V}_i,\mathbf{C}_i,\boldsymbol{\theta}^{(k)}]|\mathbf{V}_i,\mathbf{C}_i,\boldsymbol{\theta}^{(k)}\}\\
&= E\{\mathbf{y}_iE[u_iE(\mathbf{z}_i^\top|\mathbf{y}_i,u_i,\mathbf{V}_i,\mathbf{C}_i,\boldsymbol{\theta}^{(k)})|\mathbf{y}_i,\mathbf{V}_i,\mathbf{C}_i,\boldsymbol{\theta}^{(k)}]|\mathbf{V}_i,\mathbf{C}_i,\boldsymbol{\theta}^{(k)}\}\\
&= E\left\{\mathbf{y}_iE\left[u_i(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta})^\top\boldsymbol{\varphi}_i^\top|\mathbf{y}_i,\mathbf{V}_i,\mathbf{C}_i,\boldsymbol{\theta}^{(k)}\right]\left|\mathbf{V}_i,\mathbf{C}_i,\boldsymbol{\theta}^{(k)}\right.\right\}\\
&= E\left[\mathbf{y}_iE(u_i|\mathbf{y}_i,\mathbf{V}_i,\mathbf{C}_i,\boldsymbol{\theta}^{(k)})(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta})^\top\boldsymbol{\varphi}_i^\top\left|\mathbf{V}_i,\mathbf{C}_i,\boldsymbol{\theta}^{(k)}\right.\right]\\
&= E\left\{\left[\left(\frac{(\nu+p)}{(\nu+\delta)}\mathbf{y}_i\mathbf{y}_i^\top\right)-\left(\frac{(\nu+p)}{(\nu+\delta)}\mathbf{y}_i\right)\boldsymbol{\beta}^\top\mathbf{X}_i^\top\right]\boldsymbol{\varphi}_i^\top\left|\mathbf{V}_i,\mathbf{C}_i,\boldsymbol{\theta}^{(k)}\right.\right\}\\
&= \left[E\left(\frac{(\nu+p)}{(\nu+\delta)}\mathbf{y}_i\mathbf{y}_i^\top\left|\mathbf{V}_i,\mathbf{C}_i,\boldsymbol{\theta}^{(k)}\right.\right)-E\left(\frac{(\nu+p)}{(\nu+\delta)}\mathbf{y}_i\left|\mathbf{V}_i,\mathbf{C}_i,\boldsymbol{\theta}^{(k)}\right.\right)\boldsymbol{\beta}^\top\mathbf{X}_i^\top\right]\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Delta}^\top\\
&= \left[\frac{T_p(\mathbf{a}|\boldsymbol{\mu},\boldsymbol{\Sigma}^*,\nu+2)}{T_p(\mathbf{a}|\boldsymbol{\mu},\boldsymbol{\Sigma},\nu)}E\{\mathbf{W}_i\}-\frac{T_p(\mathbf{a}|\boldsymbol{\mu},\boldsymbol{\Sigma}^*,\nu+2)}{T_p(\mathbf{a}|\boldsymbol{\mu},\boldsymbol{\Sigma},\nu)}E\{\mathbf{W}_i\mathbf{W}_i^\top\}\boldsymbol{\beta}^\top\mathbf{X}_i^\top\right]\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Delta}^\top\\
&= \left[\widehat{u_i\mathbf{y}_i\mathbf{y}_i^\top}-\widehat{u_i\mathbf{y}_i}\boldsymbol{\beta}^\top\mathbf{X}_i^\top\right]\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Delta}^\top.
\end{aligned}
$$

Then, the conditional expectation of the complete log-likelihood $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})$ gives

$$
\begin{aligned}
Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}) \propto &- \frac{1}{2}\sum_{i=1}^n\left[\log|\boldsymbol{\Psi}|+\widehat{a_i}^{(k)}-2(\widehat{u_i\mathbf{y}_i}^{(k)}-\boldsymbol{\Lambda}\widehat{u_i\mathbf{z}_i}^{(k)})\boldsymbol{\beta}^\top\mathbf{X}_i^\top\boldsymbol{\Psi}^{-1}+\widehat{u}_i^{(k)}\mathbf{X}_i\boldsymbol{\beta}\boldsymbol{\beta}^\top\mathbf{X}_i^\top\boldsymbol{\Psi}^{-1}\right]\\
&- \frac{1}{2}\sum_{i=1}^n\left[\log|\boldsymbol{\Omega}|+tr\left(\boldsymbol{\Omega}^{-1}\widehat{u_i\mathbf{z}_i\mathbf{z}_i^\top}^{(k)}\right)\right],
\end{aligned}
$$

where

$$
\widehat{a_i}^{(k)} = tr\left(\widehat{u_i\mathbf{y}_i\mathbf{y}_i^\top}^{(k)}\boldsymbol{\Psi}^{-1(k)}-2\widehat{u_i\mathbf{y}_i\mathbf{z}_i^\top}^{(k)}\boldsymbol{\Lambda}^{\top(k)}\boldsymbol{\Psi}^{-1(k)}+\boldsymbol{\Lambda}^{(k)}\widehat{u_i\mathbf{z}_i\mathbf{z}_i^\top}^{(k)}\boldsymbol{\Lambda}^{\top(k)}\boldsymbol{\Psi}^{-1}\right).
$$

Taking the derivatives with respect to $\boldsymbol{\beta}$, $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$ leads to

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

where

$$
\begin{aligned}
\widehat{B}_i^{(k)} = {} & \mathrm{tr}(\widehat{u_i\mathbf{y}_i\mathbf{y}_i^\top}^{(k)})-\widehat{u_i\mathbf{y}_i}^{\top(k)}\mathbf{X}_i\boldsymbol{\beta}-\mathrm{tr}(\widehat{u_i\mathbf{y}_i\mathbf{z}_i^\top}^{(k)}\boldsymbol{\Lambda})-\boldsymbol{\beta}^\top\mathbf{X}_i^\top\widehat{u_i\mathbf{y}_i}^{(k)}+\boldsymbol{\beta}^\top\mathbf{X}_i^\top\widehat{u_i}^{(k)}\mathbf{X}_i\boldsymbol{\beta}\\
&+\boldsymbol{\beta}^\top\mathbf{X}_i^\top\boldsymbol{\Lambda}\widehat{u_i\mathbf{z}_i}^{(k)}-\mathrm{tr}(\widehat{u_i\mathbf{y}_i\mathbf{z}_i^\top}^{(k)}\boldsymbol{\Lambda}^\top)+\widehat{u_i\mathbf{z}_i}^{\top(k)}\boldsymbol{\Lambda}^\top\mathbf{X}_i\boldsymbol{\beta}+\mathrm{tr}(\widehat{u_i\mathbf{z}_i\mathbf{z}_i^\top}^{(k)}\boldsymbol{\Lambda}^\top\boldsymbol{\Lambda}).
\end{aligned}
$$

The solution of these derivatives at zero gives the estimates of the MLE presented in (8)-(10).

References

Akaike, H., 1974. A new look at the statistical model identification. IEEE Trans. Autom. Cont. 19, 716–723.

Arellano-Valle, R. B., Bolfarine, H., Lachos, V. H., 2005. Skew-normal linear mixed models. Journal of Data Science 3, 415–438.

Arellano-Valle, R. B., Genton, M. G., 2010. Multivariate extended skew-t distributions and related families. Metron 68 (3), 201–234.

Azzalini, A., Genton, M., 2008. Robust likelihood methods based on the skew-t and related distributions. International Statistical Review 76, 1490–1507.

Bozdogan, H., 1987. Model selection and akaike's information criterion (aic): The general theory and its analytical extensions. Psychometrika 52 (3), 345–370.

Costa, D. R., Lachos, V. H., Bazan, J. L., Azevedo, C. L. N., 2013. Estimation methods for multivariate tobit confirmatory factor analysis. Tech. Rep. 18, Department of Statistics, Campinas State University.

Dempster, A., Laird, N., Rubin, D., 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B, 39, 1–38.

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.

Ho, H. J., Lin, T.-I., Chen, H.-Y., Wang, W.-L., 2012. Some results on the truncated multivariate t distribution. Journal of Statistical Planning and Inference 142 (1), 25–40.

Jacqmin-Gadda, H., Thiebaut, R., Chene, G., Commenges, D., 2000. Analysis of left-censored longitudinal data with application to viral load in HIV infection. Biostatistics 1 (4), 355–368.

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

Lange, K. L., Little, R. J., Taylor, J. M., 1989. Robust statistical modeling using the t distribution. Journal of the American Statistical Association 84 (408), 881–896.

Lin, T.-I., Wu, P. H., McLachlan, G. J., Lee, S. X., 2013. The skew-t factor analysis model. arXiv preprint arXiv:1310.5336.

Louis, T., 1982. Finding the observed information matrix when using the EM algorithm. Journal of the Royal Statistical Society, Series B, 226–233.

Lucas, A., 1997. Robustness of the Student t based M-estimator. Communications in Statistics: Theory and Methods 26, 1165–1182.

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

Matos, L. A., Prates, M. O., H-Chen, M., Lachos, V., 2013b. Likelihood-based inference for mixed-effects models with censored response using the multivariate-t distribution. Statistica Sinica 23, 1323–1342.

McLachlan, G., Bean, R., Ben-Tovim Jones, L., 2007. Extension of the mixture of factor analyzers model to incorporate the multivariate *t*-distribution. Computational Statistics & Data Analysis 51, 5327–5338.

Meng, X. L., Rubin, D. B., 1993. Maximum likelihood estimation via the ECM algorithm: a general framework. Biometrika 80, 267–278.

Meza, C., Osorio, F., de la Cruz, R., 2012. Estimation in non-linear mixed effects models using heavy tailed distributions. Statistics and Computing 22, 121–139.

Muthén, B. O., 1989. Tobit factor analysis. British Journal Mathematics and Statistics Psychology 42, 241–250.

Prates, M. O., Costa, D. R., Lachos, V. H., 2013. Generalized linear mixed models for correlated binary data with t-link. Statistics and Computing, doi: 10.1007/s11222–013–9423–3.

R Development Core Team, 2009. 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`

Schwarz, G., 1978. Estimating the dimension of a model. The annals of statistics 6 (2), 461–464.

Vaida, F., Fitzgerald, A., DeGruttola, V., 2007. Efficient hybrid EM for linear and nonlinear mixed effects models with censored response. Computational Statistics & Data Analysis 51 (12), 5718–5730.

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.

Wang, W., Lin, T., 2013. An efficient ECM algorithm for maximum likelihood estimation in mixtures of t-factor analyzers. Computational Statistics 28, 751–769.

Wu, L., 2010. Mixed Effects Models for Complex Data. Chapman & Hall/CRC, Boca Raton, FL.

Zhang, J., Li, J., Liu, C., 2013. Robust factor analysis using the multivariate t-distribution. Statistica Sinica. In press.

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.

Zhou, X., Tan, C., 2010. Maximum likelihood estimation of Tobit factor analysis for multivariate t-distribution. Communications in Statistics - Simulation and Computation 39, 1–16.

APÊNDICE D

Extension to Item Response Theory model

Two-parameter Item Response Theory model with t-link distribution

1 Introduction

This is a work in progress. We have developed a new approach to IRT models with t-link distribution that it is an extension of the previous papers in this thesis. The empirical studies are still being implemented.

Item response theory (IRT) is a set of measurement models which has gained high visibility in recent years due to its large applicability. In broader terms, IRT associates, stochastically, a latent trait and item parameters. In IRT context, it is very common to assume normality of the link function, as well as of the latent trait [see: Albert (1992), Baker and Kim (2004), Bock and Aitkin (1981)]. However, in many situations, these conjectures do not fit the data properly.

Using a more general parametric class of links is one popular way of preventing misspecification of links, such as the probit and logit links. In this work, we propose a robust parametric modeling of the unidimensional two-parameter IRT model for binary data based on the symmetric t-link function. Under our proposition, the probit and logit links can be considered special cases. Due to the complexity of the likelihood function, the use of an appropriate estimation method plays a crucial role. Generalizing the approach of Azevedo and Andrade (2013), a fully likelihood-based approach is carried out, including the implementation of an exact ECM algorithm for maximum likelihood (ML) estimation. We show that the E-step reduces the problem to computing the first two moments of certain truncated multivariate-t distributions. The general formulas for these moments were derived by Ho et al. (2012) (eq. 12 and 13). They require the multivariate-t cumulative density function (cdf), for which we use the *mvtnorm* package Genz et al. (2008) in R (R Development Core Team, 2009). The likelihood function is easily computed as a by-product of the E-step and is used for monitoring convergence and for model selection, such as the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC) and the Likelihood Ratio Test (LRT). To monitor the convergence of the proposed EM, we follow Tan et al. (2007) to directly calculate the log-likelihood values and then to plot the difference of the consecutive log-likelihood values against the EM iteration.

The rest of this chapter is organized as follows: in Section 2, we introduce the two-parameter IRT model with t-link. The proposed ECM algorithm is developed in Section 3.

2 The two-parameter t-link IRT model

Consider the situation where m subjects are submitted to an instrument composed of k items. Let Y_{ij} denote the binary outcome 0 or 1 of the i-th item and $\mathbf{Y}_j = (Y_{j1}, \ldots, Y_{jn_j})^\top$ be the collection of responses from subject j, where $j = 1, \ldots, m$ and $i = 1, \ldots, n_j$ ($n_j \leq k$). The proposed unidimensional two-parameter IRT with t-link model is defined by

$$\begin{aligned} Y_{ij} = \mathbb{I}_{(0,\infty)}(Z_{ij}),\ \mathbf{Z}_j|\theta_j, U_j = u_j &\sim N_{n_j}(\boldsymbol{\mu}_j, u_j^{-1}\mathbf{I}_{n_j}), \\ \theta_j|U_j = u_j &\sim N_1(0, u_j^{-1}), \\ U_j &\sim Gamma(\nu/2, \nu/2), \end{aligned} \tag{1}$$

where $\mu_{ij} = \alpha_i\theta_j - \beta_i$, α_i is the discrimination parameter for the item i, β_i is the difficulty parameter for the item i, θ_j is the latent trait of the subject j. When the $\boldsymbol{\alpha}$ parameters are equal to 1, we have the one-parameter IRT model, as known as Rasch model (Andrade et al., 2000). Using Lemma 1, given in Appendix A, the model defined in (1) is equivalent to the following representation

$$Pr(Y_{ij} = 1|\theta_j) = T_1(\mu_{ij}|0, 1, \nu), \quad \theta_j \sim t_1(0, 1, \nu).$$

The observed-data likelihood of parameter vector $\boldsymbol{\zeta} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\theta})$ is given by

$$L(\boldsymbol{\zeta}) = \prod_{j=1}^{m} \int t_1(\theta_j|0, 1, \nu)\psi_i^t(\boldsymbol{\zeta})d\theta_j, \tag{2}$$

where $t_1(.|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$ is as defined in Chapter 1 and

$$\psi_i^t(\boldsymbol{\zeta}) = \prod_{i=1}^{n_j} [T_1(\mu_{ij}|0, 1, \nu)]^{y_{ij}}[1 - T_1(\mu_{ij}|0, 1, \nu)]^{1-y_{ij}}.$$

Note that the likelihood function (2) does not have a closed-form expression because the model function is not linear in the random effect. To compute ML estimates of the unknown variance parameters, the EM algorithm is proposed. This algorithm relies on formulas for the mean and variance of a truncated multivariate t-distribution, which can be computed using available formula.

3 The EM algorithm

In this section, we derive the M- and E- step for the proposed ECM algorithm. For the probit IRT model several algorithms have been proposed [Baker and Kim (2004), Azevedo and Andrade (2013)]. Here we will present the methodology for the two-parameter t-link model.

Let $\boldsymbol{\theta} = \{\theta_j\}_{j=1}^m$, $\mathbf{Z} = \{\mathbf{Z}_j\}_{j=1}^m$ and $\mathbf{U} = \{U_j\}_{j=1}^m$ as missing data and $\mathbf{Y} = \{\mathbf{Y}_j\}_{j=1}^m$ the observed data. From the definition of the latent variable $\mathbf{Z}$, we have $\{\mathbf{Y},\mathbf{Z}\} = \mathbf{Z}$. Then, the joint density for the complete-data $\mathbf{Y}_{com} = \{\mathbf{Y},\mathbf{Z},\boldsymbol{\theta},\mathbf{U}\}$ is

$$
\begin{aligned}
f(\mathbf{Y}_{com}|\boldsymbol{\zeta}) &= \prod_{j=1}^m f(\theta_j|u_j) f(\mathbf{Z}_j|\theta_j,u_j,\boldsymbol{\alpha},\boldsymbol{\beta}) h(u_j|\nu) \\
&= \prod_{j=1}^m \phi_1(\theta_j|0,u_j^{-1})\phi_{n_j}(\mathbf{Z}_j|\boldsymbol{\mu}_j,u_j^{-1}\mathbf{I}_{n_j}) h(u_j|\nu). \qquad (3)
\end{aligned}
$$

The M-step of the ECM algorithm is to find the complete-data MLE of $\boldsymbol{\zeta}$ by maximizing the conditional expectation of the complete-data log-likelihood $\ell(\boldsymbol{\zeta}|\mathbf{Y}_{com}) = \log f(\mathbf{Y}_{com}|\boldsymbol{\zeta})$ given the observed data $\mathbf{Y}$ and the current estimate $\boldsymbol{\zeta}^{(k)}$, given by

$$
\begin{aligned}
Q(\boldsymbol{\zeta}|\widehat{\boldsymbol{\zeta}}^{(k)}) &= C - \frac{1}{2}\sum_{j=1}^m tr\left(E[U_j\mathbf{Z}_j^\top\mathbf{Z}_j|\mathbf{Y}_j] - 2\boldsymbol{\alpha}^\top E[U_j\mathbf{Z}_j\theta_j|\mathbf{Y}_j] + 2E[U_j\mathbf{Z}_j|\mathbf{Y}_j]\boldsymbol{\beta}_j^\top\right) \\
&+ \sum_{j=1}^m tr\left(E[U_j\theta_j|\mathbf{Y}_j]\boldsymbol{\alpha}^\top\boldsymbol{\beta}_j + E[U_j\theta_j^2|\mathbf{Y}_j]\boldsymbol{\alpha}^\top\boldsymbol{\alpha} + E[U_j|\mathbf{Y}_j]\boldsymbol{\beta}_j^\top\boldsymbol{\beta}_j + E[U_j\theta_j^2|\mathbf{Y}_j]\right),
\end{aligned}
$$

where C is a constant that is independent of the parameters. Thus, in the M-step we update $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\theta}$ through the following closed form expressions:

$$
\begin{aligned}
\widehat{\boldsymbol{\alpha}} &= \left(\sum_{j=1}^m E[U_j\theta_j^2|\mathbf{Y}_j]\right)^{-1}\sum_{j=1}^m \left(E[U_j\mathbf{Z}_j\theta_j|\mathbf{Y}_j] + E[U_j\theta_j|\mathbf{Y}_j]\boldsymbol{\beta}\right), \qquad (4) \\
\widehat{\boldsymbol{\beta}} &= \left(\sum_{j=1}^m E[U_j|\mathbf{Y}_j]\right)^{-1}\sum_{j=1}^m \left(\boldsymbol{\alpha}E[U_j\theta_j|\mathbf{Y}_j] - E[U_j\mathbf{Z}_j|\mathbf{Y}_j]\right) \quad \text{and} \qquad (5) \\
\widehat{\theta}_j &= \frac{\boldsymbol{\alpha}^\top\left(E[\mathbf{Z}_j|\mathbf{Y}_j] + \boldsymbol{\beta}\right)}{1+\sum_{i=1}^k \alpha_i^2}, \qquad (6)
\end{aligned}
$$

where $E[U_j|\mathbf{Y}_j]$, $E[U_j\theta_j|\mathbf{Y}_j]$, $E[U_jZ_j|\mathbf{Y}_j]$, $E[U_j\mathbf{Z}_j\theta_j|\mathbf{Y}_j]$, $E[U_j\theta_j^2|\mathbf{Y}_j]$ and $E[\theta_j|\mathbf{Y}_j]$ are expected values in $(U_j,\theta_j,\mathbf{Z}_j)$ conditional on $\mathbf{Y}_j$, taken at the current parameters value $\boldsymbol{\zeta}^{(k)} = (\boldsymbol{\theta}^{(k)},\boldsymbol{\alpha}^{(k)},\boldsymbol{\beta}^{(k)})$.

For the E-step, we first derive the conditional predictive distribution of the missing data, which is given by:

$$
\begin{aligned}
f(\boldsymbol{\theta}, \mathbf{Z}, \mathbf{U}|\mathbf{Y}) &= f(\mathbf{Z}|\mathbf{Y}, \boldsymbol{\theta}, \mathbf{u}, \boldsymbol{\alpha}, \boldsymbol{\beta}) f(\boldsymbol{\theta}|\mathbf{Y}, \mathbf{u}, \boldsymbol{\alpha}, \boldsymbol{\beta}) f(\mathbf{u}|\mathbf{Y}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \\
&= f(\boldsymbol{\theta}|\mathbf{Y}, \mathbf{Z}, u_j, \boldsymbol{\alpha}, \boldsymbol{\beta}) f(\mathbf{u}|\mathbf{Y}, \mathbf{Z}, \boldsymbol{\alpha}, \boldsymbol{\beta}) f(\mathbf{Z}|\mathbf{Y}, \boldsymbol{\alpha}, \boldsymbol{\beta}). \qquad (7)
\end{aligned}
$$

Since $f(\boldsymbol{\theta}|\mathbf{Y}, \mathbf{Z}, \mathbf{u}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ is proportional to (3), we obtain the following result:

$$
\begin{aligned}
f(\boldsymbol{\theta}|\mathbf{Y}, \mathbf{Z}, \mathbf{u}, \boldsymbol{\alpha}, \boldsymbol{\beta}) &= \prod_{j=1}^{m} f(\theta|\mathbf{Y}_j, \mathbf{Z}_j, \mathbf{u}_j, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_{j=1}^{m} f(\theta|\mathbf{Z}_j, \mathbf{u}_j, \boldsymbol{\alpha}, \boldsymbol{\beta}) \\
&= \prod_{j=1}^{m} \phi_1 \left(\theta_j | \frac{\boldsymbol{\alpha}^\top(\mathbf{Z}_j + \boldsymbol{\beta})}{1 + \sum_{i=1}^{k} \alpha_i^2}, \frac{u_j^{-1}}{1 + \sum_{i=1}^{k} \alpha_i^2} \right),
\end{aligned}
$$

To derive the second term on the right-hand side of (7), we use the following result from Chib and Greenberg (1998):

$$
P(\mathbf{Y}_j = y_j|\theta_j, \mathbf{Z}_j, \mathbf{u}_j, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_{i=1}^{n_j} \{ I_{(Z_{ij}>0)} I_{(Y_{ij}=1)} + I_{(Z_{ij}\leq 0)} I_{(Y_{ij}=0)} \}, \qquad (8)
$$

which indicates that given $\mathbf{Z}_j$, the conditional probability of $\mathbf{Y}_j$ is independent of θ_j and $\mathbf{u}_j$. Hence, expression (8) implies $P(\mathbf{Y}_j = y_j|\mathbf{Z}_j, \boldsymbol{\alpha}, \boldsymbol{\beta}) = I_{(\mathbf{Z}_j \in B_j)}$. Since the conditional probability $Z_{ij}|u_j, \boldsymbol{\alpha}, \boldsymbol{\beta}$ is normally distributed and $U_j \sim Gamma(v/2, v/2)$, the marginal distribution of $Z_{ij}|\alpha_i, \boldsymbol{\beta}_i$ follows $t_{n_j}(-\boldsymbol{\beta}, \Sigma, \nu)$, with $\Sigma = u_j^{-1}(\mathbf{I}_{n_j} + \boldsymbol{\alpha}\boldsymbol{\alpha}^\top))$. Furthermore, from

$$
\begin{aligned}
f(\mathbf{Z}_j|\mathbf{Y}_j, \boldsymbol{\alpha}, \boldsymbol{\beta}) &\propto f(\mathbf{Z}_j, \mathbf{Y}_j|\boldsymbol{\alpha}, \boldsymbol{\beta}) \\
&= f(\mathbf{Z}_j|\boldsymbol{\alpha}, \boldsymbol{\beta}) P(\mathbf{Y}_j = \mathbf{y}_j|\mathbf{Z}_j, \boldsymbol{\alpha}, \boldsymbol{\beta}) \\
&= t_{n_j}(\mathbf{Z}_j| - \boldsymbol{\beta}, \Sigma, \nu) I_{(\mathbf{Z}_j \in B_j)}
\end{aligned}
$$

we obtain

$$
f(\mathbf{Z}|\mathbf{Y}, \boldsymbol{\beta}) = \prod_{j=1}^{m} f(\mathbf{Z}_j, \mathbf{Y}_j|\boldsymbol{\beta}) = Tt_{n_j}(\mathbf{Z}_j| - \boldsymbol{\beta}, \Sigma, \nu, B_j).
$$

Using the prior results and the property that if $\mathbf{Z}|\boldsymbol{\beta}$ follows $t_p(\mu, \sigma, \nu)$ and $U \sim Gamma(v/2, v/2)$ we have $E(U|Z) = \frac{\nu+p}{\nu+d^2}$ [see, Lachos et al. (2011)], where d^2 is the Mahalanobis distance, the estimates are:

$$
\begin{aligned}
E(U_j|\mathbf{Y}_j) &= E\left[E(U_j|\mathbf{Y}_j,\mathbf{Z}_j)|\mathbf{Y}_j\right] = E\left[\frac{\nu+n_j}{\nu+\delta_j}|\mathbf{Y}_j\right] = \bar{\mathbf{Z}}_j^0,\\
E[U_j\mathbf{Z}_j|\mathbf{Y}_j] &= E\left[\mathbf{Z}_j E(U_j|\mathbf{Y}_j,\mathbf{Z}_j)|\mathbf{Y}_j\right] = E\left[\left(\frac{\nu+n_j}{\nu+\delta_j}\right)\mathbf{Z}_j|\mathbf{Y}_j\right] = \bar{\mathbf{Z}}_j^1,\\
E[U_j\theta_j|\mathbf{Y}_j] &= E\{E\left[U_j E(\theta_j|\mathbf{Y}_j,\mathbf{Z}_j,U_j)|\mathbf{Y}_j,\mathbf{Z}_j\right]|\mathbf{Y}_j\} = \frac{\boldsymbol{\alpha}^\top}{1+\sum_{i=1}^k\alpha_i^2}(\bar{\mathbf{Z}}_j^1-\boldsymbol{\beta}\bar{\mathbf{Z}}_j^0),\\
E[\theta_j|\mathbf{Y}_j] &= E\{E\left[E(\theta_j|\mathbf{Y}_j,\mathbf{Z}_j,U_j)|\mathbf{Y}_j,\mathbf{Z}_j\right]|\mathbf{Y}_j\} = \frac{\boldsymbol{\alpha}^\top}{1+\sum_{i=1}^k\alpha_i^2}(M_1+\boldsymbol{\beta}),\\
E[U_j\mathbf{Z}_j\theta_j|\mathbf{Y}_j] &= E\{E\left[E(\theta_j|\mathbf{Y}_j,\mathbf{Z}_j,U_j)|\mathbf{Y}_j,\mathbf{Z}_j\right]|\mathbf{Y}_j\} = \frac{\boldsymbol{\alpha}^\top}{1+\sum_{i=1}^k\alpha_i^2}(\bar{\mathbf{Z}}_j^2+\bar{\mathbf{Z}}_j^1\boldsymbol{\beta}^\top),\\
E[U_j\theta_j^2|\mathbf{Y}_j] &= E\{E\left[E(\theta_j|\mathbf{Y}_j,\mathbf{Z}_j,U_j)|\mathbf{Y}_j,\mathbf{Z}_j\right]|\mathbf{Y}_j\} = \frac{1}{1+\sum_{i=1}^k\alpha_i^2}\\
&+ \frac{\boldsymbol{\alpha}^\top}{1+\sum_{i=1}^k\alpha_i^2}(\bar{\mathbf{Z}}_j^2+2\bar{\mathbf{Z}}_j^1\boldsymbol{\beta}^\top+\bar{\mathbf{Z}}_j^0\boldsymbol{\beta}\boldsymbol{\beta}^\top)\boldsymbol{\alpha},
\end{aligned}
$$

where $M_1 = E\left[\mathbf{Z}_j|\mathbf{Y}_j\right]$, $\bar{\mathbf{Z}}_j^2 = E\left[\frac{\nu+n_j}{\nu+\delta_j}\mathbf{Z}_j\mathbf{Z}_j^\top|\mathbf{Y}_j\right]$, and $\mathbb{B}_j = B_{j1}\times\ldots,\times B_{jn_j}$, where B_{ij} is the interval $(0,\infty)$ if $y_{ij}=1$ and the interval $(-\infty,0]$ if $y_{ij}=0$.

From (4)-(6), the E-step reduces to computation of M_1, $\bar{\mathbf{Z}}_i^0$, $\bar{\mathbf{Z}}_i^1$ and $\bar{\mathbf{Z}}_i^2$. It is clear that Proposition 1 of the Appendix A cannot be used, since the components of the random vector $\mathbf{Z}_j|\mathbf{Y}_j$ are right or left-truncated depending in $y_{ij}, i=1,\ldots,n_j$. However, these quantities can be determined in closed form using a sequence of simple transformations, as follows:

(i) The first step is to standardize the components of $\mathbb{B}_j$, either as left- or right-truncated. Let $\mathbf{A}_j$ be a diagonal matrix with diagonal elements equal to -1 or 1 depending on $B_{ij}=(0,\infty)$ or $B_{ij}=(-\infty,0]$, respectively. Then, $\mathbf{U}_j \equiv \mathbf{A}_j\mathbf{Z}_j|\mathbf{Y}_j \sim Tt_{n_j}(\mathbf{A}_j\boldsymbol{\gamma}_j,\ \mathbf{A}_j\boldsymbol{\Omega}_j\mathbf{A}_j,\nu;\mathbb{C}_j)$, $\mathbb{C}_j=(-\infty,0]^{n_j}$, that is, $\mathbf{U}_j$ follows a multivariate t-distribution $t_{n_j}(\mathbf{A}_j\boldsymbol{\gamma}_j,\mathbf{A}_j\boldsymbol{\Omega}_j\mathbf{A}_j,\nu)$ right truncated at $(-\infty,0]^{n_j}$. This standardization facilitates the computation of $\bar{\mathbf{U}}_j^0 = E\left[\frac{\nu+n_j}{\nu+\delta_j^u}|\mathbf{Y}_j\right]$, $\bar{\mathbf{U}}_j^1 = E\left[\frac{\nu+n_j}{\nu+\delta_j^u}\mathbf{U}_j|\mathbf{Y}_j\right]$, $\bar{\mathbf{U}}_j^2 = E\left[\frac{\nu+n_j}{\nu+\delta_j^u}\mathbf{U}_j\mathbf{U}_j^\top|\mathbf{Y}_j\right]$, through the result given in Proposition 1 along with the computation of the first two moments of a truncated multivariate-t distribution with specific parameters, where $\delta_j^u = (\mathbf{U}_j-\mathbf{A}_j\boldsymbol{\gamma}_j)^\top(\mathbf{A}_j\boldsymbol{\Omega}_j\mathbf{A}_j)^{-1}(\mathbf{U}_j-\mathbf{A}_j\boldsymbol{\gamma}_j)$.

(ii) The second step is to note that $\bar{\mathbf{Z}}_j^0 = \bar{\mathbf{U}}_j^0$, $\bar{\mathbf{Z}}_j^1 = \mathbf{A}_j^{-1}\bar{\mathbf{U}}_j^1$ and $\bar{\mathbf{Z}}_j^2 = \mathbf{A}_j^{-1}\bar{\mathbf{U}}_j^2\mathbf{A}_j^{-1}$, since $\delta_j^u = \delta_j = (\mathbf{Z}_j - \boldsymbol{\gamma}_j)^\top \boldsymbol{\Omega}_j^{-1}(\mathbf{Z}_j - \boldsymbol{\gamma}_j)$.

When ν goes to ∞, we have an interesting EM-type algorithm to the two-parameter probit IRT model.

References

Albert, J. H., 1992. Bayesian estimation of normal ogive item response curves using gibbs sampling. Journal of Educational and Behavioral Statistics 17 (3), 251–269.

Andrade, D., Tavares, H., Valle, R., 2000. Item response theory: Concepts and applications. Associação Brasileira de Estatística, São Paulo (in Portuguese).

Azevedo, C. L. N., Andrade, D. F., 2013. CADEM: A conditional augmented data EM algorithm for fitting one parameter probit models. Brazilian Journal of Probability and Statistics 27 (2), 245–262.

Baker, F. B., Kim, S.-H., 2004. Item response theory: Parameter estimation techniques. Vol. 176. CRC.

Bock, R. D., Aitkin, M., 1981. Marginal maximum likelihood estimation of item parameters: Application of an em algorithm. Psychometrika 46 (4), 443–459.

Chib, S., Greenberg, E., 1998. Analysis of multivariate probit models. Biometrika 85, 347–361.

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.

Ho, H. J., Lin, T.-I., Chen, H.-Y., Wang, W.-L., 2012. Some results on the truncated multivariate t distribution. Journal of Statistical Planning and Inference 142 (1), 25–40.

Lachos, V. H., Angolini, T., Abanto-Valle, C. A., 2011. On estimation and local influence analysis for measurement errors models under heavy-tailed distributions. Statistical Papers 52, 567–590.

R Development Core Team, 2009. 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`

Tan, M., Tian, G., Fang, H., 2007. An efficient MCEM algorithm for fitting generalized linear mixed models for correlated binary data. Journal of Statistical Computation and Simulation 77 (11), 929–943.

Referências Bibliográficas

Akaike, H., 1974. A new look at the statistical model identification. IEEE Trans. Autom. Cont. 19, 716–723.

Albert, J., Chib, S., 1993. Bayesian analysis of binary and polychotomous response data. Journal of the American Statistical Association 88, 669–679.

An, X., Bentler, P. M., 2012. Efficient direct sampling MCEM algorithm for latent variable models with binary responses. Computational Statistics & Data Analysis 56 (2), 231–244.

Bartholomew, D. J., 1984. The foundations of factor analysis. Biometrika 71 (2), 221–232.

Bartholomew, D. J., Knott, M., Moustaki, I., 2011. Latent variable models and factor analysis: A unified approach. Vol. 899. Wiley.

Bliss, C. I., 1935. The calculation of the dosage-mortality curve. Annals of Applied Biology 22 (1), 134–167.

Bozdogan, H., 1987. Model selection and akaike's information criterion (AIC): The general theory and its analytical extensions. Psychometrika 52 (3), 345–370.

Breslow, N., Clayton, D., 1993. Approximate inference in generalized linear mixed models. Journal of the American Statistical Association 88, 9–25.

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.

Czado, C., Santner, T., 1992. The effect of link misspecification on binary regression inference. Journal of Statistical Planning and Inference 33, 213–231.

Delyon, B., Lavielle, M., Moulines, E., 1999. Convergence of a stochastic approximation version of the EM algorithm. Annals of Statistics 27 (1), 94–128.

Dempster, A., Laird, N., Rubin, D., 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B, 39, 1–38.

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.

Fernandez, C., Steel, M. F., 1999. Multivariate student-t regression models: Pitfalls and inference. Biometrika 86 (1), 153–167.

Geman, S., Geman, D., 1984. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. Pattern Analysis and Machine Intelligence, IEEE Transactions on (6), 721–741.

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.

Greene, W. H., 2005. Censored data and truncated distributions. Available at SSRN 825845.

Halekoh, U., Højsgaard, S., Yan, J., 2006. The r package geepack for generalized estimating equations. Journal of Statistical Software 15 (2), 1–11.

Ho, H. J., Lin, T.-I., Chen, H.-Y., Wang, W.-L., 2012. Some results on the truncated multivariate t distribution. Journal of Statistical Planning and Inference 142 (1), 25–40.

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.

Jamshidian, M., 1999. Adaptive robust regression by using a nonlinear regression program. Journal of Statistical Software 4, 1–25.

Johnson, S., Narasimhan, B., 2011. Package cubature. R package version 1.1-1, URL http://cran.r-project.org/web/packages/cubature/index.html.

Jöreskog, 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.

Lachos, V. H., Cabral, C. R., Garay, A. M., 2013a. Modelos não lineares assimétricos. Associação Brasileira de Estatística, São Paulo.

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

Lange, K. L., Little, R. J., Taylor, J. M., 1989. Robust statistical modeling using the t distribution. Journal of the American Statistical Association 84 (408), 881–896.

Lange, K. L., Sinsheimer, J. S., 1993. Normal/independent distributions and their applications in robust regression. Journal of Computational and Graphical Statistics 2, 175–198.

Lawley, D. N., Maxwell, A. E., 1971. Factor analysis as a statistical method. Vol. 18. Butterworths London.

Lee, Y., Nelder, J., 2006. Double hierarchical generalized linear models. Applied Statistics 55 (Part 2), 139–185.

Li, J., Liu, C., Zhang, J., 2007. Robust factor analysis using the multivariate t-distribution. Metrica 29, 178–181.

Liu, C., 2004. Robit regression: a simple robust alternative to logistic and probit regression. Applied Bayesian modeling and causal inference from incomplete-data perspectives, 227–238.

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

Louis, T., 1982. Finding the observed information matrix when using the EM algorithm. Journal of the Royal Statistical Society, Series B, 226–233.

Lucas, A., 1997. Robustness of the Student t based M-estimator. Communications in Statistics: Theory and Methods 26, 1165–1182.

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

Matos, L. A., Prates, M. O., H-Chen, M., Lachos, V., 2013b. Likelihood-based inference for mixed-effects models with censored response using the multivariate-t distribution. Statistica Sinica 23, 1323–1342.

McCulloch, C., 1994. Maximum likelihood variance components estimation for binary data. Journal of the American Statistical Association 89 (425), 330–335.

McCulloch, C. E., 1997. Maximum Likelihood Algorithms for Generalized Linear Mixed Models. Journal of the American statistical Association 92, 162–170.

McLachlan, G., Krishnan, T., 1997. The EM algorithm and extensions. Wiley New York.

Meng, X., van Dyk, D., 1998. Fast EM-type implementations for mixed effects models. Journal of the Royal Statistical Society. Series B, Statistical Methodology 60 (3), 559–578.

Meza, C., Jaffrezic, F., Foulley, J., 2009. Estimation in the probit normal model for binary outcomes using the SAEM algorithm. Computational Statistics & Data Analysis 53 (4), 1350–1360.

Meza, C., Osorio, F., de la Cruz, R., 2012. Estimation in non-linear mixed effects models using heavy tailed distributions. Statistics and Computing 22, 121–139.

Mulaik, S. A., 1972. The foundations of factor analysis. McGraw-Hill New York.

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.

Pinheiro, J. C., Liu, C. H., Wu, Y. N., 2001. Efficient algorithms for robust estimation in linear mixed-effects models using a multivariate t-distribution. Journal of Computational and Graphical Statistics 10, 249–276.

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/`

Robert, C., Casella, G., Robert, C., 1999. Monte Carlo Statistical Methods. Vol. 2. Springer New York.

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

Schwarz, G., 1978. Estimating the dimension of a model. The Annals of Statistics 6 (2), 461–464.

Skrondal, A., Rabe-Hesketh, S., 2004. Generalized latent variable modeling. Chapman & Hall/CRC.

Spearman, C., 1904. "general intelligence,"objectively determined and measured. The American Journal of Psychology 15 (2), 201–292.

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

Tan, M., Tian, G., Fang, H., 2007. An efficient MCEM algorithm for fitting generalized linear mixed models for correlated binary data. Journal of Statistical Computation and Simulation 77 (11), 929–943.

Tan, M. T., Tian, G.-L., Ng, K. W., 2010. Bayesian missing data problems: EM, data augmentation and noniterative computation. Vol. 32. Chapman and Hall/CRC.

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

Vaida, F., Fitzgerald, A., DeGruttola, V., 2007. Efficient hybrid EM for linear and nonlinear mixed effects models with censored response. Computational Statistics & Data Analysis 51 (12), 5718–5730.

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., Muthén, 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.

Wu, L., 2010. Mixed Effects Models for Complex Data. Chapman & Hall/CRC, Boca Raton, FL.

Zeller, C. B., 2009. Distribuições misturas de escala skew-normal: estimação e diagnostico em modelos lineares.

Zhang, J., Li, J., Liu, C., 2013. Robust factor analysis using the multivariate t-distribution. Statistica Sinica In press.

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.

Zhou, X., Tan, C., 2010. Maximum likelihood estimation of Tobit factor analysis for multivariate t-distribution. Communications in Statistics - Simulation and Computation 39, 1–16.