

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

Denise Reis Costa^a

Marcos O. Prates^b

Víctor H. Lachos^{a,*}

^a*Departamento de Estatística, Universidade Estadual de Campinas, Brazil*

^b*Departamento de Estatística, Universidade Federal de Minas Gerais, Brazil*

Abstract

A critical issue in modelling binary response data is the choice of the links. We introduce a new link based on the Student t-distribution (t-link) for correlated binary data. The t-link relates to the common probit-normal link adding one additional parameter which purely 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 rely on available formulas for the mean and variance of a truncated multivariate t-distribution. To illustrate the new methodology, 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 & 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 & van Dyk, 1998), maximum likelihood (ML) for fitting GLMM for correlated binary responses remains to be 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

*Corresponding author. Address for correspondence: Departamento de Estatística, Rua Sérgio Buarque de Holanda, 651, Cidade Universitária Zeferino Vaz, Campinas, São Paulo, Brazil. CEP 13083-859. e-mail: hlachos@ime.unicamp.br

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) propose to use the Stochastic Approximation version of EM (SAEM), proposed by Delyon *et al.* (1999), to obtain the ML estimates. Lee & Nelder (2006) proposed an approximated procedure based on the h -likelihood. However, by its 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 & 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 modeling of GLMM for binary data based on the symmetric t-link so that the t-GLMM is defined. A fully 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 link 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 Lin *et al.* (2011) (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 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-t 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

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

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

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

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 cdf will be denoted by $T_p(\cdot | \boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$. 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

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

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 $f(\mathbf{x} | \boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu; \mathbb{A}) = \frac{t_p(\mathbf{x} | \boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)}{T_p(\mathbf{a} | \boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)} \mathbb{I}_{\mathbb{A}}(\mathbf{x})$, 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 elsewhere. The following result is related with 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 Appendix.

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

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

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

where $c_p(\nu, r) = \left(\frac{\nu + p}{\nu} \right)^r \left(\frac{\Gamma((p + \nu)/2) \Gamma((\nu + 2r)/2)}{\Gamma(\nu/2) \Gamma((p + \nu + 2r)/2)} \right)$, $\delta = (\mathbf{X} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{X} - \boldsymbol{\mu})$, $\mathbf{a} = (a_1, \dots, a_p)^\top$, $\boldsymbol{\Sigma}^* = \frac{\nu}{\nu + 2r} \boldsymbol{\Sigma}$, $\mathbf{X}^{(0)} = 1$, $\mathbf{X}^{(1)} = \mathbf{X}$, $\mathbf{X}^{(2)} = \mathbf{X}\mathbf{X}^\top$ 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 Lin *et al.* (2011), 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 j th measurement and $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{in_i})^\top$ be the collection of responses from subject i , where $i = 1, \dots, m$ and $j = 1, \dots, n_i$. 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

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

where $\Phi(\cdot)$ denotes the cumulative distribution function of $N(0, 1)$, $\mathbf{x}_{ij}^\top = (x_{ij1}, \dots, x_{ijp})^\top$ and $\mathbf{w}_{ij}^\top = (w_{ij1}, \dots, w_{ijq})^\top$ 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 $\mathbf{Z}_i = (Z_{i1}, \dots, Z_{in_i})^\top$, such that

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

$$\mathbf{b}_i \sim N_q(\mathbf{0}, \mathbf{D}),$$

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

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

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

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

where $\phi_q(\cdot | \boldsymbol{\mu}, \boldsymbol{\Sigma})$ 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 $\mathbf{y}_i = (y_{i1}, \dots, y_{in_i})^\top$ denotes a realization of \mathbf{Y}_i .

Under normality, several methods have been proposed to compute efficiently 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 & Nelder (2006). Later, Tan *et al.* (2007) proposed a non-iterative importance sampling approach (IBF) where the first and the second order moments of a truncated multivariate normal distribution were evaluated and associated with the MCEM algorithm. Another approach, proposed to use SAEM algorithm (Delyon *et al.*, 1999) which 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 & Chib, 1993). In the next section we present the robust t-link for GLMM.

3.2 The t-GLMM for binary data

The proposed model is defined by

$$Y_{ij} = \mathbb{I}_{(0,\infty)}(Z_{ij}), \quad \mathbf{Z}_i | \mathbf{b}_i, U_i = u_i \sim N_{ni}(\boldsymbol{\mu}_i, u_i^{-1} \mathbf{I}_{n_i}),$$

$$\mathbf{b}_i | U_i = u_i \sim N_q(\mathbf{0}, u_i^{-1} \mathbf{D}),$$

$$U_i \sim \text{Gamma}(v/2, v/2).$$
(3)

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

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

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

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

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

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

$$\begin{aligned} \psi_i^t(\boldsymbol{\beta}, \mathbf{b}_i) = & \prod_{j=1}^{n_i} [T_1(\mu_{ij} | 0, 1, \boldsymbol{\nu})]^{y_{ij}} \times \\ & [1 - T_1(\mu_{ij} | 0, 1, \boldsymbol{\nu})]^{1-y_{ij}}. \end{aligned} \quad (5)$$

Note that the likelihood (5) 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.

On the other hand, although several authors have addressed parameter estimation of mixing variables in model regressions (see, for example, Lange & Sinsheimer (1993); Jamshidian (1999)) based upon the EM algorithm and its variants, S. (1999) discusses potential problems that may arise in the estimation of degrees of freedom, in particular for the Student- 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 & Sinsheimer (1993) are questionable because they do not provide sufficient information on whether these estimates correspond to local, or global maximas. Interestingly, Lucas (1997) notes that only under fixed degrees of freedom, the parameter estimates behave robust against extreme observations. A plausible (and simple) alternative is to assume that the parameter $\boldsymbol{\nu}$ 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 $\boldsymbol{\nu}$ to be known. See for instance, Meza *et al.* (2009) and Vanegas & Cysneiros (2010).

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, let 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 $\mathbf{Y} = \{\mathbf{Y}_i\}_{i=1}^m$ the observed data. Then, the joint density for the complete-data $\mathbf{Y}_{com} = \{\mathbf{Y}, \mathbf{Z}, \mathbf{b}, \mathbf{U}\}$ is

$$\begin{aligned} L(\boldsymbol{\theta} | \mathbf{Y}_{com}) = & \prod_{i=1}^m [\phi_q(\mathbf{b}_i | \mathbf{0}, u_i^{-1} \mathbf{D}) \phi_{n_i}(\mathbf{Z}_i | \mu_i, u_i^{-1} \mathbf{I}_{n_i}) \\ & \times h(u_i | \boldsymbol{\nu})], \end{aligned}$$

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

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

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

$$\begin{aligned}
\hat{\beta} &= \left(\sum_{i=1}^m E[U_i | \mathbf{Y}_i] \mathbf{X}_i^\top \mathbf{X}_i \right)^{-1} \sum_{i=1}^m (\mathbf{X}_i^\top E[U_i \mathbf{Z}_i | \mathbf{Y}_i] - \\
&\quad - \mathbf{W}_i E[U_i \mathbf{b}_i | \mathbf{Y}_i]) \quad \text{and} \quad (6)
\end{aligned}$$

$$\hat{\mathbf{D}} = \frac{1}{m} \sum_{i=1}^m E[U_i \mathbf{b}_i \mathbf{b}_i^\top | \mathbf{Y}_i], \quad (7)$$

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 parameters value $\theta^{(k)} = (\beta^{(k)}, \mathbf{D}^{(k)})$. It can be show that (see Appendix)

$$\begin{aligned}
E[U_i | \mathbf{Y}_i] &= \bar{\mathbf{Z}}_i^0, \quad E[U_i \mathbf{Z}_i | \mathbf{Y}_i] = \bar{\mathbf{Z}}_i^1, \\
E[U_i \mathbf{b}_i | \mathbf{Y}_i] &= \Delta_i (\bar{\mathbf{Z}}_i^1 - \bar{\mathbf{Z}}_i^0 \mathbf{X}_i \beta), \\
E[U_i \mathbf{b}_i \mathbf{b}_i^\top | \mathbf{Y}_i] &= \Lambda_i + \Delta_i (\bar{\mathbf{Z}}_i^2 + \gamma_i \gamma_i^\top \bar{\mathbf{Z}}_i^0 - \bar{\mathbf{Z}}_i^1 \gamma_i^\top - \\
&\quad - \gamma_i \bar{\mathbf{Z}}_i^{1\top}) \Delta_i, \\
\mathbf{Z}_i | \mathbf{Y}_i &\sim t_{n_i}(\gamma_i, \Omega_i, \nu) \mathbb{I}_{\mathbb{B}_i}(\mathbf{Z}_i), \quad (8)
\end{aligned}$$

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

From (6)-(7), the E-step reduces to computation of $\bar{\mathbf{Z}}_i^0$, $\bar{\mathbf{Z}}_i^1$ and $\bar{\mathbf{Z}}_i^2$. From (8) 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 in y_{ij} , $j = 1, \dots, n_i$. 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, $\mathbf{U}_i \equiv \mathbf{A}_i \mathbf{Z}_i | \mathbf{Y}_i \sim T t_{n_i}(\mathbf{A}_i \gamma_i, \mathbf{A}_i \Omega_i \mathbf{A}_i, \nu; \mathbb{C}_i)$, $\mathbb{C}_i = (-\infty, 0]^{n_i}$, that is, \mathbf{U}_i follows a multivariate t-distribution $t_{n_i}(\mathbf{A}_i \gamma_i, \mathbf{A}_i \Omega_i \mathbf{A}_i, \nu)$ right truncated at $(-\infty, 0]^{n_i}$. This standardization facilitate 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 $\delta_i^u = (\mathbf{U}_i - \mathbf{A}_i \gamma_i)^\top (\mathbf{A}_i \Omega_i \mathbf{A}_i)^{-1} (\mathbf{U}_i - \mathbf{A}_i \gamma_i)$.

- (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 $\delta_i^u = \delta_i = (\mathbf{Z}_i - \gamma_i)^\top \Omega_i^{-1}(\mathbf{Z}_i - \gamma_i)$.

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

4.1 Estimation of the likelihood

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

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

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

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

where $\mathbf{b}_1, \dots, \mathbf{b}_l, \dots, \mathbf{b}_K$ are draws from $\tilde{\pi}(\mathbf{b}_i, \theta)$. An efficient choice for $\tilde{\pi}$ consists of the conditional distribution of \mathbf{b}_i given the data \mathbf{y}_i Robert *et al.* (1999).

Another approach to calculate the likelihood is to consider the *adaptIntegrate* function at *cubeature* package (Johnson & N., 2011) available in R. This function is appropriate for numerical multivariate integration and will be used to calculate the log-likelihood for 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 $\ell(\theta)$, like $\|\ell(\hat{\theta}^{(k+1)}) - \ell(\hat{\theta}^{(k)})\|$ or $\|\ell(\hat{\theta}^{(k+1)})/\ell(\hat{\theta}^{(k)}) - 1\|$, $k = 0, 2, \dots$, is small enough. In addition, model selections based on the observed likelihood can be done using appropriate likelihood ratios test, AIC and BIC.

4.2 Standard error approximation

Denote the ML estimates from the EM-algorithm by $\hat{\theta} = (\hat{\beta}, \hat{\mathbf{D}})$. Under some general regularity conditions, we follow McLachlan & 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 $I_o(\hat{\theta} | \mathbf{Y}) = \sum_{i=1}^m \hat{\mathbf{u}}_i \hat{\mathbf{u}}_i^\top$ to be the observed information matrix, where $\hat{\mathbf{u}}_i = E \left[\frac{\partial \ell_i(\theta | \mathbf{Y}_{comp})}{\partial \theta} | \mathbf{y}_i \right] \Big|_{\hat{\theta}}$ with $\ell_i(\theta | \mathbf{Y}_{comp})$ being the complete-data log-likelihood formed from the single observation \mathbf{y}_i , $i = 1, \dots, m$.

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

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

where 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 $I_o(\hat{\theta} | \mathbf{Y})$.

5 Applications

5.1 Simulation study

A simulation is presented to study the performance of our proposed methodology. 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 to verify if the comparison measures AIC and BIC determines the corrected model to the simulated data. The simulation study mimics the structure of the real data analysis presented in Section 5.2.

We consider the following framework for our generalized linear mixed model:

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

where $\boldsymbol{\mu}_i = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 X_{i1} + \boldsymbol{\beta}_2 X_{i2}$ with $\boldsymbol{\beta} = (1, -0.3, 0.8)^\top$, X_1 is a continuous covariate, X_2 is a dichotomous covariate and $F(\boldsymbol{\nu})$ is a positive distribution.

To fulfill our objectives we present three simulation scenarios:

1. probit: Data has a probit link and normal random effects. So, $U = 1$ is a degenerated distribution with probability of $P(U = 1) = 1$;
2. t-link: Data has a Student-t link and Student-t random effects with degrees of freedom $\nu = 4$. So, $U \sim \text{Gamma}(\nu/2, \nu/2)$;
3. contaminated-link: Data has a contaminated normal link and contaminated normal random effects with degrees of freedom $\boldsymbol{\nu} = (\nu_1, \nu_2)^\top = (0.1, 0.1)^\top$. So, $U = \begin{cases} \nu_2 & \text{with prob } \nu_1 \\ 1 & \text{with prob } 1 - \nu_1 \end{cases}$.

For all scenarios, sample sizes of 100, 250 and 500 were generated. Under the three scenarios we fitted the probit-GLMM (Section 3.1) and the t-GLMM with 4 degrees of freedom (Section 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 250 simulated datasets under each setting. For each simulation, the parameters estimation as well as AIC and BIC were recorded.

Table 1 presents the summary statistics for $\boldsymbol{\beta}$ (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 250 estimates given by $\sum_{j=1}^{250} \hat{\gamma}_j / 250$ and MC Sd is the arithmetic average of the 250 standard deviations of the estimates of the parameters given by $\sum_{j=1}^{250} sd(\hat{\gamma}_j) / 250$, where $\gamma = \beta_0, \beta_1$ or β_2 . In addition, we also estimate the MC coverage of β_0, β_1 and β_2 , 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 as well as the t-GLMM were able to recover the true parameters values and the estimation improves as the sample size increases. Moreover, Table 1 also provides the average values of the approximate standard deviations of the EM estimates obtained through the information-based method, described in Subsection 4.2 (IM

Table 1: Results based on 250 simulated probit samples. MC mean, MC Sd (in parenthesis) and MC Coverage are the respective mean estimates, standard deviations and coverage proportion average from fitting the probit-GLMM and the t-GLMM. IM Sd are the average values of the approximate standard errors obtained through the information-based method. MC AIC and MC BIC are the arithmetic average of the respective model comparison measures.

Sample Size	Fit	Simulated probit data			MC AIC	MC BIC			
		β_0	β_1	β_2					
100	probit	MC Mean	0.981	-0.324	0.889	556.55	572.19		
		IM Sd	0.181	0.152	0.382				
		MC Sd	(0.201)	(0.154)	(0.478)				
		MC Coverage	97%	96%	98%				
	t-link	MC Mean	1.035	-0.354	0.986			558.52	574.15
		IM Sd	0.189	0.160	0.401				
		MC Sd	(0.215)	(0.171)	(0.616)				
		MC Coverage	98%	97%	99%				
250	probit	MC Mean	1.010	-0.338	0.830	1367.12	1388.25		
		IM Sd	0.125	0.096	0.261				
		MC Sd	(0.128)	(0.098)	(0.285)				
		MC Coverage	98%	97%	96%				
	t-link	MC Mean	1.064	-0.370	0.910			1371.05	1392.18
		IM Sd	0.134	0.102	0.290				
		MC Sd	(0.136)	(0.109)	(0.369)				
		MC Coverage	96%	95%	99%				
500	probit	MC Mean	1.006	-0.345	0.824	2729.10	2754.39		
		IM Sd	0.103	0.079	0.177				
		MC Sd	(0.090)	(0.068)	(0.195)				
		MC Coverage	90%	88%	96%				
	t-link	MC Mean	1.060	-0.376	0.908			2737.24	2762.52
		IM Sd	0.111	0.087	0.202				
		MC Sd	(0.096)	(0.076)	(0.253)				
		MC Coverage	88%	81%	98%				

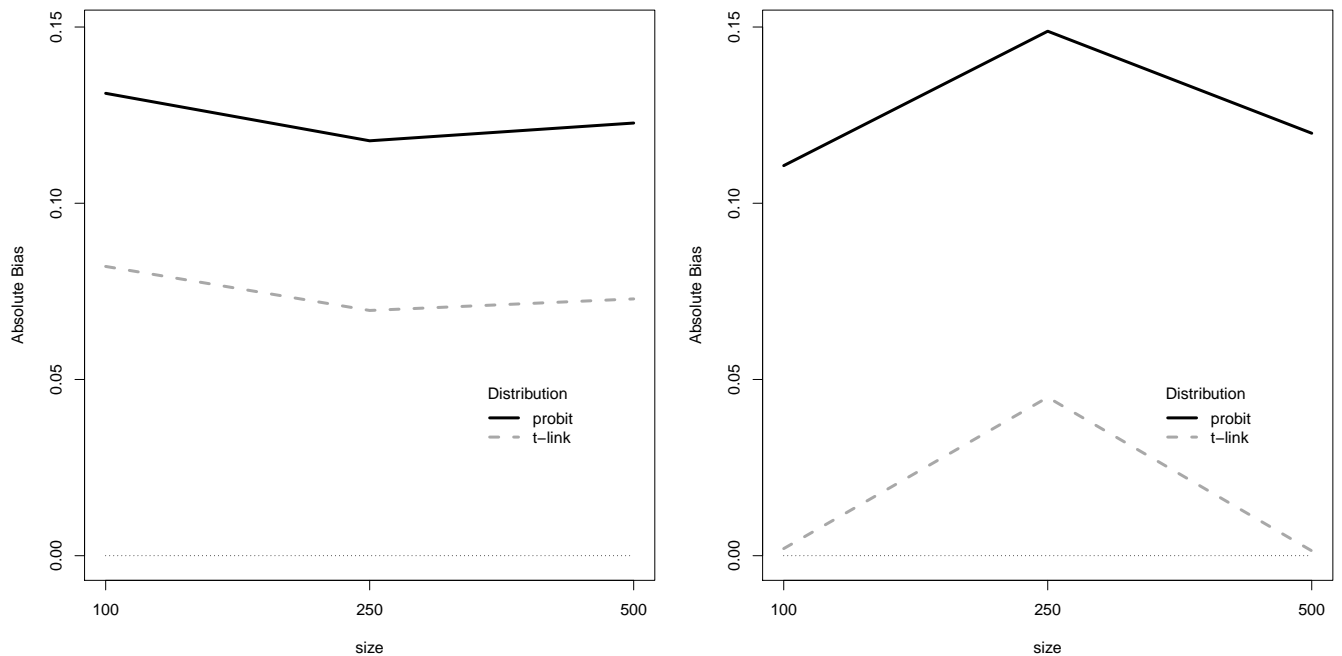


Figure 1: Simulation study. Comparison between the probit-GLMM and t-GLMM under the scenario 3. (Left panel) represents the bias of β_1 . (Right panel) represents the bias of β_3 .

Sd), and the Monte Carlo standard deviation (MC Sd) for the parameters. As we can see, the theoretically estimated standard deviation are relatively close to the Monte Carlos standard deviation estimates obtained empirically. This results show 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 average (MC AIC and MC BIC) of the model comparison criterions mentioned earlier. As we can see, the MC AIC and MC BIC correctly selects 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. From Table 2 we can see that the t-GLMM fits better the data as expected, since the true generating model was a t-GLMM. However, when the data is not generated from a probit, the probit-GLMM do 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. Equivalently to the results obtained in Table 1, we can conclude that the Mc Sd is adequately calculated. Similarly to what occurred with the probit scenario the MC AIC and MC BIC still selects the correct generating model.

Finally, we introduce scenario 3 where the main objective is to study how the probit-GLMM and the t-GLMM behaves when neither of them are the true generating model. From Table 3 we can see that the t-GLMM adjust 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 selects it as the preferred model. Therefore, the t-link is more robust to deviations on the model assumptions fitting better than the probit-GLMM when neither is the true generating model. As seen in the previous scenarios the IM Sd estimates are close to the MC Sd.

Table 2: Results based on 250 simulated t-link samples. MC mean, MC Sd (in parenthesis) and MC Coverage are the respective mean estimates, standard deviations and coverage proportion average from fitting probit-GLMM and the t-GLMM models. IM Sd are the average values of the approximate standard errors obtained through the information-based method. MC AIC and MC BIC are the arithmetic average of the respective model comparison measures.

Sample Size	Fit		Simulated t-link data			MC AIC	MC BIC
			β_0	β_1	β_2		
100	probit	MC Mean	0.946	-0.305	0.766	575.10	572.31
		IM Sd	0.177	0.152	0.414		
		MC Sd	(0.205)	(0.157)	(0.465)		
		MC Coverage	96%	96%	97%		
	t-link	MC Mean	0.996	-0.332	0.859	572.91	570.12
		IM Sd	0.187	0.161	0.446		
		MC Sd	(0.219)	(0.173)	(0.582)		
		MC Coverage	98%	97%	98%		
250	probit	MC Mean	0.968	-0.308	0.715	1420.83	1418.04
		IM Sd	0.121	0.097	0.243		
		MC Sd	(0.130)	(0.096)	(0.277)		
		MC Coverage	95%	96%	96%		
	t-link	MC Mean	1.017	-0.334	0.802	1415.98	1413.19
		IM Sd	0.129	0.103	0.262		
		MC Sd	(0.139)	(0.106)	(0.349)		
		MC Coverage	97%	95%	98%		
500	probit	MC Mean	0.944	-0.309	0.716	2849.67	2846.88
		IM Sd	0.102	0.089	0.176		
		MC Sd	(0.092)	(0.068)	(0.191)		
		MC Coverage	88%	88%	96%		
	t-link	MC Mean	0.993	-0.334	0.813	2840.14	2837.35
		IM Sd	0.115	0.100	0.200		
		MC Sd	(0.098)	(0.075)	(0.239)		
		MC Coverage	92%	86%	98%		

Table 3: Results based on 250 simulated probit-contaminated normal samples. MC mean, MC Sd (in parenthesis) and MC Coverage are the respective mean estimates, standard deviations and coverage proportion average from fitting probit and the t-link models. IM Sd are the average values of the approximate standard errors obtained through the information-based method. MC AIC and MC BIC are the arithmetic average of the respective model comparison measures.

Sample Size	Fit		Simulated contaminated-link data			MC AIC	MC BIC
			β_0	β_1	β_2		
100	probit	MC Mean	0.869	-0.286	0.711	590.96	588.17
		IM Sd	0.179	0.129	0.358		
		MC Sd	(0.191)	(0.147)	(0.431)		
		MC Coverage	90%	98%	98%		
	t-link	MC Mean	0.918	-0.311	0.798	588.58	585.79
		IM Sd	0.188	0.136	0.386		
		MC Sd	(0.205)	(0.162)	(0.536)		
		MC Coverage	95%	98%	99%		
250	probit	MC Mean	0.897	-0.287	0.681	1458.25	1457.29
		IM Sd	0.121	0.102	0.260		
		MC Sd	(0.121)	(0.092)	(0.256)		
		MC Coverage	84%	94%	90%		
	t-link	MC Mean	0.947	-0.312	0.775	1452.60	1451.64
		IM Sd	0.130	0.110	0.277		
		MC Sd	(0.131)	(0.102)	(0.324)		
		MC Coverage	92%	95%	96%		
500	probit	MC Mean	0.877	-0.295	0.704	2918.17	2918.60
		IM Sd	0.089	0.076	0.168		
		MC Sd	(0.086)	(0.065)	(0.180)		
		MC Coverage	70%	93%	93%		
	t-link	MC Mean	0.927	-0.322	0.801	2908.69	2909.12
		IM Sd	0.099	0.083	0.197		
		MC Sd	(0.092)	(0.072)	(0.227)		
		MC Coverage	82%	91%	98%		

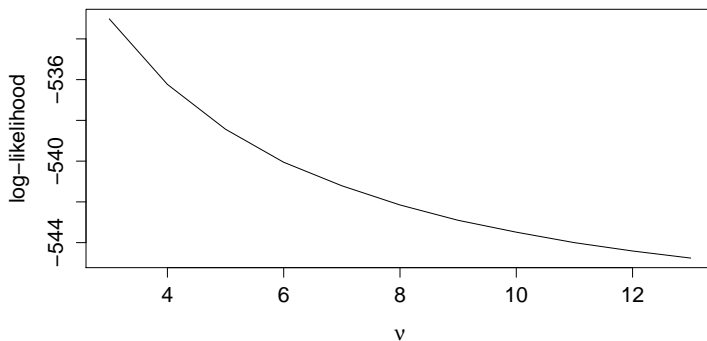


Figure 2: Plot of the profile log-likelihood ranging the degrees of freedom, v .

5.2 Analysis of real data

In this section, we analyzed the dataset presented by Skrondal & Rabe-Hesketh (2004). The data consist of a cohort study of respiratory infection in 275 Indonesian preschool children examined up to six consecutive quarters. Zeger & Karim (1991) and Skrondal & Rabe-Hesketh (2004) estimate a logistic-normal random intercept model for this data set. Here, we assume that the probability $Pr(Y_{ij} = 1 | \mathbf{b}_i)$ of respiratory infection of the i -th child at the j -th observation is assumed to be equal to:

$$F(\beta_1 + \text{age} \beta_2 + \text{xero} \beta_3 + \text{gender} \beta_4 + \text{cosine} \beta_5 + \text{sine} \beta_6 + \text{height} \beta_7 + \text{stunder} \beta_8 + w_{ij}^\top b_i),$$

where “F” is the probit or t-link; “age” is the age of the child centered around 36 months; “xero = 1” for presence of xerophthalmia (an ocular manifestation of chronic vitamin A deficiency), 0 otherwise; “cosine” and “sine” represent, respectively, the cosine and sine terms of the annual cycle to capture seasonality; “gender = 1” for female gender and 0 for male; “height” is an indicator of height for age as percent of the National Center for Health Statistics (NCHS) standard (centered at 90%), which indicates lower nutritional status; “stunted = 1” for stunting, defined as being below 85% in height for age. The elements of the matrix \mathbf{W} are $w_{ij} = (1, \text{age})$, implying that both intercept and age effects are children-specific and the vector $\mathbf{b}_i | \mathbf{D} \sim N_2(0, \mathbf{D})$, $i = 1, \dots, 275$. As the data is unbalanced, the index j can vary from 1 to 6.

In Figure 2, we plot the profile 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 $v = 3$ for our analysis.

Table 4 shows the ML estimates where the probit and t-link models were considered. At the E-step of the EM algorithm, we fixed $\beta^{(0)} = (0, 0, 0, 0, 0, 0, 0, 0)^\top$ and $\mathbf{D}^{(0)} = \mathbf{I}_2$ as the initial values. We can see that the maximum likelihood estimates for both models are close. Respiratory infection is related to intercept, age, season and height. Based on AIC and BIC criteria presented on Table 4, the t-link model is clearly the preferred one.

Given the models, the comparison between the probabilities of a children having respiratory infection with relation to intercept and age estimates are presented in Figure 3. The empirical proportion (observed proportion of having respiratory infection at each level of the variable age) is also shown. We can observe that the greater is the age of the children, the lower the propensity

Table 4: ML estimates results for fitting the respiratory infection data with probit and t-link models.

Variable	Parameter	probit		t-link	
		MLE	SE	MLE	SE
intercept	β_1	-1.594	0.143	-1.768	0.190
age	β_2	-0.018	0.006	-0.020	0.008
xero	β_3	0.352	0.300	0.327	0.436
cos(time)	β_4	-0.315	0.091	-0.436	0.135
sin(time)	β_5	-0.085	0.115	-0.103	0.159
gender	β_6	-0.302	0.186	-0.333	0.262
height	β_7	-0.026	0.016	-0.047	0.025
stunted	β_8	0.128	0.286	0.023	0.380
covariance	d_{11}	0.294	0.199	0.185	0.236
	d_{12}	-0.001	0.012	0.000	0.033
	d_{22}	0.001	0.001	0.001	0.001
AIC		1513.486		1088.036	
BIC		1553.270		1092.868	

is to have a respiratory infection. As expected, the probabilities for the t-link model are more conservative and fit the data best.

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 4. From this, we can observe that both models have succeeded in the convergence of the EM-type algorithm and we can also highlight that the t-link model was the one that presents the best result for this dataset.

6 Conclusions

In this article we have 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 benefice, the EM likelihood sequence is monotonic and the difficulties in assessing convergence which face MCMC algorithms are avoided. We conjecture that this paper introduces a novel methodology 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 the authors upon request.

Acknowledgements

V. H. Lachos and M. O. Prates would like to acknowledge the support of the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq-Brazil). V. H. Lachos would also like to

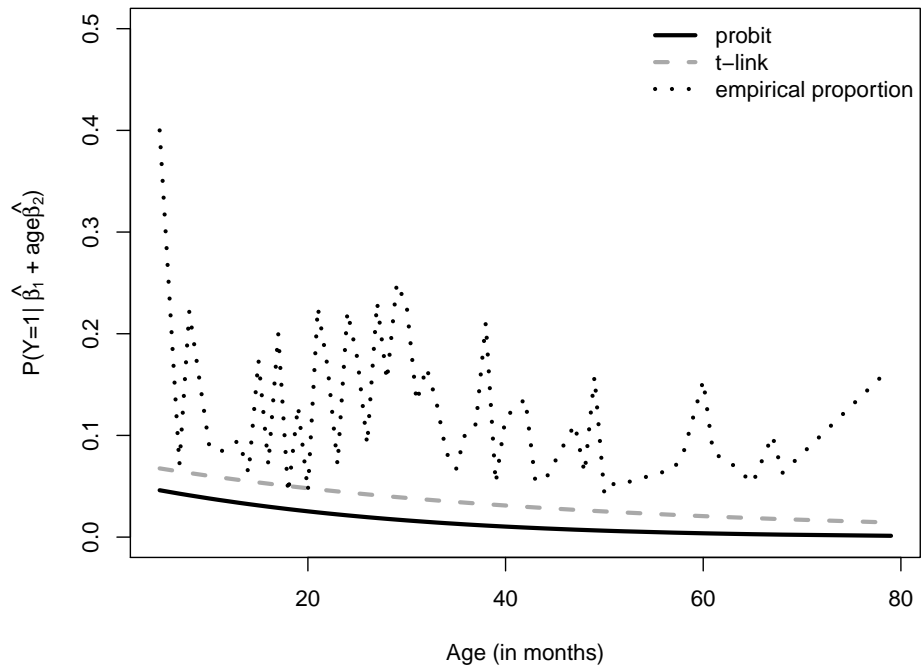


Figure 3: Probabilities of a children having respiratory infection with relation to intercept and age estimates.

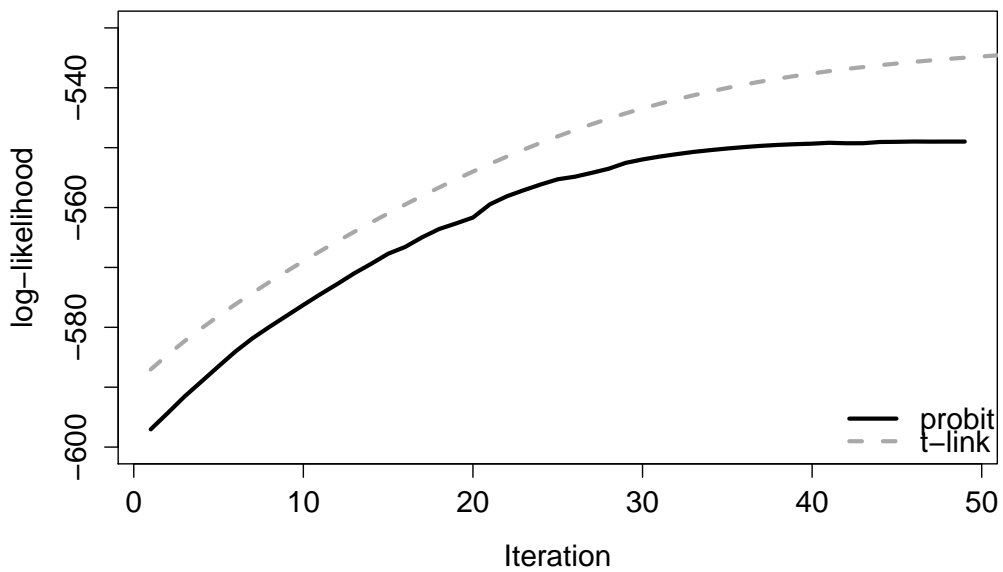


Figure 4: The comparison of convergence between two EM-exact algorithms by plotting the log-likelihood values against the EM iteration

acknowledge the support of the Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP-Brazil) and M. O. Prates would like to acknowledge the partial support of Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG-Brazil).

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

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

which concludes the proof.

Lemma 1. If $U \sim \text{Gamma}(\alpha, \beta)$, then for any vector $\mathbf{B} \in \mathbb{R}^p$ and a $p \times p$ positive definite matrix $\boldsymbol{\Sigma}$,

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

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

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

where, clearly $\mathbf{T} = \frac{\mathbf{V}}{(U\boldsymbol{\beta}/\alpha)^{1/2}}$ has a multivariate Student-t distribution, which concludes the proof. \square

Details of the EM Algorithm:

Let 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 $\{\mathbf{Y}_{obs}, \mathbf{Z}\} = \mathbf{Z}$. Then, the joint density for the complete-data $\mathbf{Y}_{com} = \{\mathbf{Y}, \mathbf{Z}, \mathbf{b}, \mathbf{U}\}$ is

$$\begin{aligned} f(\mathbf{Y}_{com} | \boldsymbol{\theta}) &= \prod_{i=1}^m f(\mathbf{b}_i | u_i, \mathbf{D}) f(\mathbf{Z}_i | \mathbf{b}_i, u_i, \boldsymbol{\beta}) h(u_i | \nu) \\ &= \prod_{i=1}^m \phi_q(\mathbf{b}_i | \mathbf{0}, u_i^{-1} \mathbf{D}) \phi_{n_i}(\mathbf{Z}_i | \boldsymbol{\mu}_i, u_i^{-1} \mathbf{I}_{n_i}) \times \\ &\quad \times h(u_i | \nu). \end{aligned} \tag{9}$$

To complete the demonstration about how to employ the EM-type algorithm for ML estimation of the t-link GLMM model, 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:

$$\begin{aligned} f(\mathbf{b}, \mathbf{Z}, \mathbf{U}|\mathbf{Y}_{obs}, \theta) &= f(\mathbf{Z}|\mathbf{Y}_{obs}, \mathbf{b}, \mathbf{u}, \theta)f(\mathbf{b}|\mathbf{Y}_{obs}, \mathbf{u}, \theta) \times \\ &\times f(\mathbf{u}|\mathbf{Y}_{obs}, \theta) = f(\mathbf{b}|\mathbf{Y}_{obs}, \mathbf{Z}, \mathbf{u}_i, \mathbf{D}) \times \\ &\times f(\mathbf{u}|\mathbf{Y}_{obs}, \mathbf{Z}, \theta)f(\mathbf{Z}|\mathbf{Y}_{obs}, \theta). \end{aligned} \quad (10)$$

Since $f(\mathbf{b}|\mathbf{Y}_{obs}, \mathbf{Z}, \mathbf{u}, \theta)$ is proportional to (9), we obtain the following result:

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

where $\Delta_i = \mathbf{D}\mathbf{W}_i^\top \Omega_i^{-1}$, $\Lambda_i = \mathbf{D} - \mathbf{D}\mathbf{W}_i^\top \Omega_i^{-1} \mathbf{W}_i \mathbf{D}$ and $\Omega_i = \mathbf{W}_i \mathbf{D}\mathbf{W}_i^\top + \mathbf{I}_{n_i}$, $i = 1, \dots, m$. To derive the second term on the right-hand side of (10), we use the following result from Chib & Greenberg (1998)

$$\begin{aligned} P(\mathbf{Y}_i = y_i|\mathbf{b}_i, \mathbf{Z}_i, \mathbf{u}_i, \theta) &= I_{(\mathbf{Z}_i \in B_i)} \\ &= \prod_{j=1}^{n_i} \{I_{(Z_{ij} > 0)}I_{(Y_{ij}=1)} + I_{(Z_{ij} \leq 0)}I_{(Y_{ij}=0)}\}, \end{aligned} \quad (11)$$

which indicates that given \mathbf{Z}_i , the conditional probability of \mathbf{Y}_i is independent of \mathbf{b}_i and \mathbf{u}_i . Hence, expression (11) implies $P(\mathbf{Y}_i = y_i|\mathbf{Z}_i, \theta) = I_{(\mathbf{Z}_i \in B_i)}$. Since the conditional probability $Z_i|u_i, \theta$ is normally distributed and $U_i \sim \text{Gamma}(v/2, v/2)$, the marginal distribution of $Z_i|\theta$ follows $t_{n_i}(X_i^\top \boldsymbol{\beta}, \Omega_i, v)$. Furthermore, from

$$\begin{aligned} f(\mathbf{Z}_i|\mathbf{Y}_i, \theta) &\propto f(\mathbf{Z}_i, \mathbf{Y}_i|\theta) \\ &= f(\mathbf{Z}_i|\theta)P(\mathbf{Y}_i = y_i|\mathbf{Z}_i, \theta) \\ &= t_{n_i}(\mathbf{Z}_i|X_i^\top \boldsymbol{\beta}, \Omega_i, v)I_{(\mathbf{Z}_i \in B_i)} \end{aligned}$$

we obtain

$$\begin{aligned} f(\mathbf{Z}|\mathbf{Y}_{obs}, \theta) &= \prod_{j=1}^m f(\mathbf{Z}_j, \mathbf{Y}_j|\theta) \\ &= T t_{n_i}(\mathbf{Z}_i|X_i^\top \boldsymbol{\beta}, \Omega_i, v, B_i). \end{aligned}$$

Using the prior results and the property that, if $Z|\theta$ follows $t_p(\boldsymbol{\mu}, \boldsymbol{\sigma}, v)$ and $U \sim \text{Gamma}(v/2, v/2)$ we have $E(U|Z) = \frac{v+p}{v+\delta}$ (see Lachos *et al.* (2011)), where δ is the Mahalanobis distance, the esti-

mates are:

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

where $\bar{\mathbf{Z}}_i^2 = E\left[\frac{\nu + n_i}{\nu + \delta_i}\mathbf{Z}_i\mathbf{Z}_i^\top|\mathbf{Y}_i\right]$, $\delta_i = (\mathbf{Z}_i - \gamma_i)^\top\Omega_i^{-1}(\mathbf{Z}_i - \gamma_i)$, $\Delta_i = \mathbf{D}\mathbf{W}_i^\top\Omega_i^{-1}$, $\Lambda_i = \mathbf{D} - \mathbf{D}\mathbf{W}_i^\top\Omega_i^{-1}\mathbf{W}_i\mathbf{D}$, $\Omega_i = \mathbf{W}_i\mathbf{D}\mathbf{W}_i^\top + \mathbf{I}_{n_i}$, $\gamma_i = \mathbf{X}_i\boldsymbol{\beta}$, and $\mathbb{B}_i = B_{i1} \times \dots \times B_{in_i}$, 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**(422), 669–679.
- Breslow, N. & Clayton, D. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, **88**(421), 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**(2), 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.
- Diggle, P. J., Kenward, P. J., Liang, K. Y. & Zeger, S. L. (2002). *Analysis of Longitudinal Data*. Oxford University Press.
- 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>.
- Jamshidian, M. (1999). Adaptive robust regression by using a nonlinear regression program. *Journal of Statistical Software*, **4**, 1–25.

- Johnson, S. G. & N., 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. *J. Comput. Graph. Stat*, **2**, 175–198.
- Lee, Y. & Nelder, J. (2006). Double hierarchical generalized linear models. *Applied Statistics*, **55**(Part 2), 139–185.
- Lin, T., Ho, H., Chen, H. & Wang, W. (2011). Some results on the truncated multivariate t distribution. *Journal of Statistical Planning and Inference*.
- Liu, C. (2004). Robit regression: a simple robust alternative to logistic and probit regression. *Applied Bayesian modeling and causal inference from incomplete-data perspectives*, pages 227–238.
- Lucas, A. (1997). Robustness of the student t based M-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*, page To appear.
- McCulloch, C. (1994). Maximum likelihood variance components estimation for binary data. *Journal of the American Statistical Association*, **89**(425), 330–335.
- McCulloch, C. (1997). Maximum Likelihood Algorithms for Generalized Linear Mixed Models. *Journal of the American statistical Association*, **92**(437).
- 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., Jaffrézic, 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.
- 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.
- Robert, C., Casella, G. & Robert, C. (1999). *Monte Carlo statistical methods*, volume 2. Springer New York.
- S., F. C. M. (1999). Multivariate student-t regression models: Pitfalls and inference. *Biometrika*, **86**, 153–168.
- Skrondal, A. & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling*. Chapman Hall/CRC.

- 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.
- Vanegas, L. & Cysneiros, F. (2010). Assesment of diagnostic procedures in symmetrical nonlinear regression models. *Computational Statistics & Data Analysis*, **54**, 1002–1016.
- Zeger, S. L. & Karim, M. R. (1991). Generalized linear models with random effects: A Gibbs sampling approach. *Journal of the American Statistical Association*, **86**, 79–86.