Likelihood Based Inference for Linear and Nonlinear Mixed-Effects Models with Censored Response Using the Multivariate-t Distribution

Larissa A. Matos

Universidade Estadual de Campinas, Brazil

Marcos O. Prates

Universidade Estadual de Campinas, Brazil

Ming H. Chen

University of Connecticut, U.S.A

Victor H. Lachos†

Universidade Estadual de Campinas, Brazil

Summary. Mixed models are commonly used to represent longitudinal or repeated measures data. An additional complication arises when the response is censored, for example, due to limits of quantification of the assay used. Normal distributions for random effects and residual errors are usually assumed, but such assumptions make inferences vulnerable to the presence of outliers. Motivated by a concern of sensitivity to potential outliers or data with tails longer-than-normal, we aim to develop a likelihood based inference for linear and nonlinear mixed effects models with censored response (NLMEC/LMEC) based on the multivariate Student-t distribution, being a flexible alternative to the use of the corresponding normal distribution. We propose an ECM algorithm for computing the maximum likelihood estimates for NLMEC/LMEC with standard errors of the fixed effects and likelihood function as a by-product. This algorithm uses closed-form expressions at the E-step, which relies on formulas for the mean and variance of a truncated multivariate-t distribution, and can be computed using available software. The proposed algorithm is implemented in the R package tlmec. An appendix which includes further mathematical details, the R code, and datasets for examples and simulations are available as supplements. The newly developed procedures are illustrated with two case studies, involving the analysis of longitudinal HIV viral load in two recent AIDS studies. In addition, a simulation study is conducted to assess the performance of the proposed approach and its comparison with the approach by Vaida and Liu (2009).

†*Address for correspondence*: Depto Estatística, IMECC-UNICAMP, Cidade Universitaria Zeferino Vaz, Distrito de Barão Geraldo, CEP 13083-859, Campinas, SP, Brazil. Fax: +55 19 3521 6078. E-mail: hlachos@ime.unicamp.br

Keywords: Censored data; HIV viral load; ECM Algorithm; Influential observations; Linear mixed models; Outliers.

1. Introduction

Linear and nonlinear mixed effects models (LME/NLME) are frequently used to analyze grouped data because they model flexibly the within-subject correlation often present in this type of data (Pinheiro and Bates, 2000). Examples of grouped data include longitudinal data, repeated measures, and multilevel data. However, in many longitudinal studies, such as studies on environmental pollution and infection diseases, measurement of some variables may be subjects to a detention limit, i.e., a certain threshold value below or above which the measurement are not quantifiable. For instance, viral load measures the amount of actively replicating virus and depending upon the diagnostic assays used, its measurement may be subjected to some upper and lower detection limits (hence, left or right censored), below or above which they are not quantifiable. The proportion of censored data in these studies may not be trivial and considering crude/adhoc methods, namely, substituting threshold value or some arbitrary point such as midpoint between zero and cutoff for detection (Vaida and Liu, 2009) might lead to biased estimates of fixed effects and variance components (Wu, 2010). As alternatives to crude imputation methods, Hughes (1999) proposed a likelihood-based Monte Carlo expectation-maximization (MCEM) algorithm for LME with censored responses (LMEC). Vaida et al. (2007) proposed a hybrid EM (HEM) algorithm for linear and nonlinear mixed effects models with censored response (LMEC/NLMEC) using a more efficient implementation of Hughes algorithm based on an efficient block-sampling scheme. Vaida and Liu (2009) proposed an exact EM-type algorithm for LMEC/NLMEC which uses closed-form expressions at the E-step, as opposed to Monte Carlo Simulation, leading to an improvement in the speed of computation of up to an order of magnitude. More recently, Matos et al. (2011) provided some additional tools, including influence diagnostics analyses for LMEC/NLMEC.

In the framework of LMEC/NLMEC, the random effects and the within-subject errors are routinely assumed to have a normal distribution for mathematical convenience. However, such normality assumptions may not always be realistic because they are vulnerable to the presence of atypical observations. To deal with the problem of atypical observations in LME with complete responses, some proposals have been made in the literature by replacing the assumption of normality by a more flexible class of distributions. For instance, Pinheiro et al. (2001) proposed a multivariate-t linear mixed model (t-LME) and demonstrated its robustness against outliers through an application to orthodontic data and extensive simulations. Lin and Lee (2007) developed some additional tools for t-LME from a Bayesian perspective. Rosa et al. (2003) advocate the use of a subclass of elliptical distributions, called normal/independent (NI) distributions (Liu, 1996) and adopted a Bayesian framework to carry out posterior analysis for heavy–tailed LME (NI-LME). Further elaborations in t-LME have been studied by Song et al. (2007) and Wang and Fan (2011). More recently, in the context of heavy–tailed LMEC/NLMEC, Lachos et al. (2011) advocate the use of the NI class of distributions and adopted a Bayesian framework to carry out posterior analysis. Even though, some works with elliptical distributions has recently appeared in the literature, there are no studies on censored LMEC/NLMEC under the Student-t family from a frequentist perspective. In this paper we propose a robust parametric modeling of LMEC/NLMEC based on the multivatiate-t distribution so that the t-LMEC/t-NLMEC is defined and a fully likelihood based approach is considered, including the implementation of an exact ECM algorithm for maximum likelihood (ML) estimation. As in Vaida and Liu (2009), 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 (LR).

The rest of the paper is organized as follows. In Section 2, we establish notation and outline some main results related with the the multivariate-t and truncated-t distribution. In Section 3 the t-LMEC and related likelihood based inference is presented. In sections 4 and 5 the extension to more general t-LMEC and to t-NLMEC, respectively, is discussed. The advantage of the proposed methodology is illustrated through the analysis of two case studies of modelling HIV viral load in Section 6. Section 7 presents a simulation study to compare the performance of our proposed methods with other normality based methods. Section 8 concludes with some discussions and possible directions for future research.

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

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

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

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

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 $\boldsymbol{\Sigma}$. The family of t-distributions thus provides a heavy-tailed alternative to the normal family with mean $\boldsymbol{\mu}$ and covariance matrix that is equal to a scalar multiple of $\boldsymbol{\Sigma}$ (if $\nu > 2$). In order to introduce some notation, for a Student-t random vector, we establish the following Proposition which is important for our subsequent research.

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

i) $\mathbf{Y}_1 \sim t_{p_1}(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_{11},\nu)$,

ii) The conditional cdf of $\mathbf{Y}_2|\mathbf{Y}_1 = \mathbf{y}_1$ *is given by*

$$P(\mathbf{Y}_2 \le \mathbf{y}_2|\mathbf{Y}_1 = \mathbf{y}_1) = T_{p_2}\left(\mathbf{y}_2|\boldsymbol{\mu}_{2.1},\widetilde{\boldsymbol{\Sigma}}_{22.1},\nu+p_1\right), \tag{2}$$

i.e., $\mathbf{Y}_2|\mathbf{Y}_1 = \mathbf{y}_1 \sim t_{p_2}\left(\boldsymbol{\mu}_{2.1},\widetilde{\boldsymbol{\Sigma}}_{22.1},\nu+p_1\right)$, *where* $\widetilde{\boldsymbol{\Sigma}}_{22.1} = \left(\frac{\nu+\delta_1}{\nu+p_1}\right)\boldsymbol{\Sigma}_{22.1}$, $\delta_1 = (\mathbf{y}_1-\boldsymbol{\mu}_1)^\top\boldsymbol{\Sigma}_{11}^{-1}(\mathbf{y}_1-\boldsymbol{\mu}_1)$, $\boldsymbol{\Sigma}_{22.1} = \boldsymbol{\Sigma}_{22}-\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}$, $\boldsymbol{\mu}_{2.1} = \boldsymbol{\mu}_2+\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\mathbf{y}_1-\boldsymbol{\mu}_1)$, *and* $T_p(.|\boldsymbol{\mu},\boldsymbol{\Sigma},\nu)$ *denotes the cdf of the* p*-variate Student-t distribution with parameters* $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ *and* ν.

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

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

$$\mathbb{A} = \{\mathbf{x} = (x_1,\ldots,x_p)^\top|x_1 \le a_1,\ldots,x_p \le a_p\}. \tag{3}$$

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:

$$f(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu; \mathbb{A}) = \frac{t_p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)}{T_p(\mathbf{a}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)} \mathbb{I}_{\mathbb{A}}(\mathbf{x}), \tag{4}$$

where $\mathbf{a} = (a_1, \ldots, a_p)^\top$ 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 propositions, are crucial for evaluating some conditional expectations of the proposed ECM algorithm for censored mixed effects models.

PROPOSITION 2. *If* $\mathbf{X} \sim Tt_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu; \mathbb{A})$, *with* $\mathbb{A}$ *as defined in (3), then*

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

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

PROOF. First note that if $\mathbf{X} \sim t_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$, then we can write

$$\left(\frac{\nu+p}{\nu+\delta}\right)^r t_p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu) = c_p(\nu, r) t_p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}^*, \nu+2r). \tag{5}$$

It follows that

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

which concludes the proof.

PROPOSITION 3. *Let* $\mathbf{X} \sim Tt_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu; \mathbb{A})$, *with* $\mathbb{A}$ *as defined in (3). Consider the partition* $\mathbf{X}^\top = (\mathbf{X}_1^\top, \mathbf{X}_2^\top)$ *with* $dim(\mathbf{X}_1) = p_1$, $dim(\mathbf{X}_2) = p_2$, $p_1 + p_2 = p$, *and the corresponding partition of the parameters* $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, $\mathbf{a}$ $(\mathbf{a}^{x_1}, \mathbf{a}^{x_2})$ *and* $\mathbb{A}$ $(\mathbb{A}^{x_1}, \mathbb{A}^{x_2})$. *Then under the notation given in Proposition 1 we have*

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

where $d_p(p_1, \nu, r) = (\nu+p)^r \left(\frac{\Gamma((p+\nu)/2)\Gamma((p_1+\nu+2r)/2)}{\Gamma((p_1+\nu)/2)\Gamma((p+\nu+2r)/2)}\right)$, $\mathbf{W} \sim Tt_{p_2}(\boldsymbol{\mu}_{2.1}, \widetilde{\boldsymbol{\Sigma}}^*_{22.1}, \nu + p_1 + 2r; \mathbb{A}^{x_2})$, $\delta = (\mathbf{X}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{X}-\boldsymbol{\mu})$, $\delta_1 = (\mathbf{X}_1-\boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}_{11}^{-1}(\mathbf{X}_1-\boldsymbol{\mu}_1)$, $\mathbf{a}^{x_2} = (a_1, \ldots, a_{p2})^\top$, $\widetilde{\boldsymbol{\Sigma}}^*_{22.1} = \left(\dfrac{\nu+\delta_1}{\nu+2r+p_1}\right)\boldsymbol{\Sigma}_{22.1}$, $\mathbf{V}^{(0)} = 1$, $\mathbf{V}^{(1)} = \mathbf{V}$ $\mathbf{V}^{(2)} = \mathbf{V}\mathbf{V}^\top$ *and* $\nu + p_1 + 2r > 0$.

PROOF. 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

$$\left(\frac{\nu+p}{\nu+\delta}\right)^r t_{p_2}\left(\mathbf{x}_2|\boldsymbol{\mu}_{2.1}, \widetilde{\boldsymbol{\Sigma}}_{22.1}, \nu+p_1\right) = \frac{d_p(p_1,\nu,r)}{(\nu+\delta_1)^r} t_{p_2}(\mathbf{x_2}|\boldsymbol{\mu}_{2.1}, \widetilde{\boldsymbol{\Sigma}}^*_{22.1}, \nu+p_1+2r) \quad (6)$$

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

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) (eq. 12 and 13), which depending on the multivariate-t cdf. The computation uses existing functions for the multivariate-t cumulative distribution, for which the *pmvt*() of the *mvtnorm* library (Genz et al., 2008) from *R* can be used. A computer code to calculate the first two moments of a truncated multivariate-t distributions, written in R, is available from the first author upon request.

3. Linear mixed effects with censored response

3.1. Model specification

For robust estimation of the parameters, we proceed as in Pinheiro et al. (2001) by considering a generalization of the classical N–LME as follows:

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i, \quad (7)$$

with the assumption that

$$\begin{pmatrix} \mathbf{b}_i \\ \boldsymbol{\epsilon}_i \end{pmatrix} \overset{ind.}{\sim} t_{n_i+q}\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \mathbf{D} & \mathbf{0} \\ \mathbf{0} & \sigma^2\mathbf{I}_{n_i} \end{pmatrix}, \nu\right), i = 1, \ldots, n, \quad (8)$$

where the subscript i is the subject index; $\mathbf{I}_p$ denotes the $p \times p$ identity matrix; $\mathbf{y}_i = (Y_{i1}, \ldots, Y_{in_i})^\top$ is a $n_i \times 1$ vector of observed continuous responses for sample unit i, $\mathbf{X}_i$ is the $n_i \times p$ design matrix corresponding to the fixed effects, $\boldsymbol{\beta}$ is a $p \times 1$ vector of population-averaged regression coefficients called fixed effects, $\mathbf{Z}_i$ is the $n_i \times q$ design matrix corresponding to the $q \times 1$ vector of random effects $\mathbf{b}_i$, $\boldsymbol{\epsilon}_i$ is the $n_i \times 1$ vector of random errors, and the dispersion matrix $\mathbf{D} = \mathbf{D}(\boldsymbol{\alpha})$ depends on unknown and reduced parameters $\boldsymbol{\alpha}$.

From (8), it is clear that marginally

$$\mathbf{b}_i \stackrel{iid}{\sim} t_q(\mathbf{0}, \mathbf{D}, \nu) \quad \text{and} \quad \boldsymbol{\epsilon}_i \stackrel{iid}{\sim} t_{n_i}(\mathbf{0}, \sigma^2\mathbf{I}_{n_i}, \nu), \quad i = 1, \ldots, n. \tag{9}$$

Note that $\mathbf{b}_i$ and $\boldsymbol{\epsilon}_i$ are uncorrelated, once $Cov(\mathbf{b}_i, \boldsymbol{\epsilon}_i) = E\{\mathbf{b}_i\boldsymbol{\epsilon}_i^\top\} = E\{E\{\mathbf{b}_i\boldsymbol{\epsilon}_i^\top|U_i\}\} = \mathbf{0}$. Classical inference on the parameter vector $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \sigma^2, \boldsymbol{\alpha}^\top, \nu)^\top$ is based on the marginal distribution for $\mathbf{y}_i$, which are marginally distributed as

$$\mathbf{y}_i \stackrel{\text{ind.}}{\sim} t_{n_i}(\mathbf{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma}_i, \nu), \tag{10}$$

for $i = 1, \ldots, n$, where $\boldsymbol{\Sigma}_i = \sigma^2\mathbf{I}_{ni} + \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^\top$. The estimates from the multivariate t-LME are more robust against outliers than those based on the standard LME. In a simulation study, Pinheiro et al. (2001) showed that the t-LME substantially outperforms the normal or standard LME when outliers are present in the data. The gains in efficiency in estimating the parameter is particularly high for the variance - covariance parameters. This problem has been also discussed by Wu (2010) in the context of censored mixed effects models.

Following Vaida and Liu (2009), in this paper we consider the case in which the response Y_{ij} is not fully observed for all i, j. Thus, let the observed data for the i-th subject be $(\mathbf{Q}_i, \mathbf{C}_i)$, where $\mathbf{Q}_i$ represents the vector of uncensored reading or censoring level, and $\mathbf{C}_i$ the vector of censoring indicators:

$$y_{ij} \quad \leq \quad Q_{ij} \quad \text{if} \quad C_{ij} = 1, \text{ and } \ y_{ij} = Q_{ij} \quad \text{if} \quad C_{ij} = 0, \tag{11}$$

so that, the t-LMEC is defined. For simplicity we will assume that the data are left-censored. The extensions to arbitrary censoring are immediate. It follows that for responses with censoring pattern as in (11), we have that marginally $\mathbf{y}_i \sim Tt_{n_i}(\mathbf{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma}_i, \nu; \mathbb{A}_i)$, where $\mathbb{A}_i = A_{i1} \times \ldots, \times A_{ini}$, with A_{ij} as the interval $(-\infty, \infty)$ if $C_{ij} = 0$ and the interval $(-\infty, Q_{ij}]$ if $C_{ij} = 1$. In the next section, we present the likelihood function, which can be easily computed by using a sequence of simple steps.

3.2. The likelihood function

The first step is to treat separately the observed and censored components of $\mathbf{y}_i$. Partition $\mathbf{y}_i$ into the observed and censored parts: $\mathbf{y}_i = vec(\mathbf{y}_i^o, \mathbf{y}_i^c)$, that is, $C_{ij} = 0$ for all elements in $\mathbf{y}_i^o$, and 1 for all elements in $\mathbf{y}_i^c$; write accordingly $\mathbf{Q}_i = vec(\mathbf{Q}_i^o, \mathbf{Q}_i^c)$, where $vec(.)$ denote the function which stacks vectors or matrices of the same number of columns. Then, from

Proposition 1, we have that $\mathbf{y}_i^o \sim t_{n_i^o}(\mathbf{X}_i^o\boldsymbol{\beta}, \boldsymbol{\Sigma}_i^{oo}, \nu)$, $\mathbf{y}_i^c|\mathbf{y}_i^o, \sim t_{n_i^c}(\boldsymbol{\mu}_i^{co}, \mathbf{S}_i^{co}, \nu + n_i^o)$, where

$$\begin{aligned} \boldsymbol{\mu}_i^{co} &= \mathbf{X}_i^c\boldsymbol{\beta} + \boldsymbol{\Sigma}_i^{co}\boldsymbol{\Sigma}_i^{oo-1}(\mathbf{y}_i^o - \mathbf{X}_i^o\boldsymbol{\beta}), &(12)\\ \mathbf{S}_i^{co} &= \left(\frac{\nu + Q(\mathbf{y}_i^o)}{\nu + n_i^o}\right)\boldsymbol{\Sigma}_i^{cc.o}, &(13)\end{aligned}$$

with $\boldsymbol{\Sigma}_i^{cc.o} = \boldsymbol{\Sigma}_i^{cc} - \boldsymbol{\Sigma}_i^{co}\boldsymbol{\Sigma}_i^{oo-1}\boldsymbol{\Sigma}_i^{oc}$ and $Q(\mathbf{y}_i^o) = (\mathbf{y}_i^o - \mathbf{X}_i^o\boldsymbol{\beta})^\top\boldsymbol{\Sigma}_i^{oo-1}(\mathbf{y}_i^o - \mathbf{X}_i^o\boldsymbol{\beta})$. Thus, the likelihood for cluster i is given by

$$\begin{aligned} L_i(\boldsymbol{\theta}|\mathbf{y}) &= P(\mathbf{Q}_i|\mathbf{C}_i, \boldsymbol{\theta}) = P(\mathbf{y}_i^c \leq \mathbf{Q}_i^c|\mathbf{y}_i^o = \mathbf{Q}_i^o, \boldsymbol{\theta})P(\mathbf{y}_i^o = \mathbf{Q}_i^o|\boldsymbol{\theta}), &(14)\\ &= t_{n_i^o}(\mathbf{Q}_i^o|\mathbf{X}_i^o\boldsymbol{\beta}, \boldsymbol{\Sigma}_i^{oo}, \nu)T_{n_i^c}(\mathbf{Q}_i^c|\boldsymbol{\mu}_i^{co}, \mathbf{S}_i^{co}, \nu + n_i^o) = L_i. &(15)\end{aligned}$$

Therefore, the log-likelihood function for the observed data is given by $\ell(\boldsymbol{\theta}|\mathbf{y}) = \sum_{i=1}^n\{\log L_i\}$. This can be computed at each step of the EM-type algorithm without additional computational burden, because L_i's are computed at the E-step (see Subsection 3.3). In addition, The log-likelihood can be used to monitor the convergence of the algorithm and for model selection (AIC, BIC, LR).

Lucas (1997) developed an interesting study on 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. Otherwise, if the degrees of freedom is also estimated by maximum likelihood, the influence functions for ν and the change of variance function of the location parameter are not bounded. In this work we will maintain fixed the degrees of freedom and the shape parameters for Student-t, and we will use a model selection procedure based on the AIC or BIC to choose the most appropriate values of ν (see Lange et al., 1989; Meza et al., 2011). Thus, hereafter we consider that the parameter vector is $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \sigma^2, \boldsymbol{\alpha}^\top)^\top$.

3.3. The EM algorithm

The EM algorithm originally proposed by Dempster, Laird and Rubin (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 (7)-(8) and (11) is complicated such that the EM algorithm is less advisable due to a computational difficulty in the M-step. To cope with this problem, we apply an extension of EM algorithm, called the ECM (Meng and Rubin, 1993) algorithm, which shares the appealing features of the EM and has a typically faster convergence rate than the EM in the sense of a small amount of iterations or actual computer time.

Let $\mathbf{y} = (\mathbf{y}_1^\top, \ldots, \mathbf{y}_n^\top)^\top$, $\mathbf{b} = (\mathbf{b}_1^\top, \ldots, \mathbf{b}_n^\top)^\top$, $\mathbf{u} = (u_1, \ldots, u_n)^\top$, $\mathbf{Q} = vec(\mathbf{Q}_1, \ldots, \mathbf{Q}_n)$ and $\mathbf{C} = vec(\mathbf{C}_1, \ldots, \mathbf{C}_n)$, such that we observe $(\mathbf{Q}_i, \mathbf{C}_i)$ for the i-th subject. Treating $\mathbf{b}$, $\mathbf{u}$ and $\mathbf{y}$ as hypothetical missing data, and augmented with the observed data $\mathbf{Q}, \mathbf{C}$, we set $\mathbf{y}_c = (\mathbf{C}^\top, \mathbf{Q}^\top, \mathbf{y}^\top, \mathbf{b}^\top, \mathbf{u}^\top)^\top$. Hence, the ECM algorithm is applied to the complete-data log-likelihood function $\ell_c(\boldsymbol{\theta}|\mathbf{y}_c) = \sum_{i=1}^n \ell_i(\boldsymbol{\theta}|\mathbf{y}_c)$, given by

$$\begin{aligned} \ell_i(\boldsymbol{\theta}|\mathbf{y}_c) \quad = \quad & -\frac{1}{2}\Big[n_i \log \sigma^2 + \frac{u_i}{\sigma^2}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{Z}_i\mathbf{b}_i)^\top(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{Z}_i\mathbf{b}_i) \\ & + \log|\mathbf{D}| + u_i\mathbf{b}_i^\top\mathbf{D}^{-1}\mathbf{b}_i\Big] + h(u_i|\nu) + C, \end{aligned} \tag{16}$$

where C is a constant that is independent of the parameter vector $\boldsymbol{\theta}$ and $h(u_i|\nu)$ is a density of a $Gamma(\nu/2, \nu/2)$. 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 (see appendix)

$$Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k)}) = E[\ell_c(\boldsymbol{\theta}|\mathbf{y}_c)|\mathbf{Q}, \mathbf{C}, \widehat{\boldsymbol{\theta}}^{(k)}] = \sum_{i=1}^n Q_i(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k)}) = \sum_{i=1}^n Q_{1i}(\boldsymbol{\beta}, \sigma^2|\widehat{\boldsymbol{\theta}}^{(k)}) + \sum_{i=1}^n Q_{2i}(\boldsymbol{\alpha}|\widehat{\boldsymbol{\theta}}^{(k)}), \tag{17}$$

where

$$\begin{aligned} Q_{1i}(\boldsymbol{\beta}, \sigma^2|\widehat{\boldsymbol{\theta}}^{(k)}) \quad = \quad & -\frac{n_i}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\Big[\widehat{a}_i^{(k)} - 2\widehat{\boldsymbol{\beta}}^{(k)\top}\mathbf{X}_i^\top(\widehat{u\mathbf{y}}_i^{(k)} - \mathbf{Z}_i\widehat{u\mathbf{b}}_i^{(k)}) \\ & + \widehat{u}_i^{(k)}\widehat{\boldsymbol{\beta}}^{(k)\top}\mathbf{X}_i^\top\mathbf{X}_i\widehat{\boldsymbol{\beta}}^{(k)}\Big] \end{aligned}$$

and

$$Q_{2i}(\boldsymbol{\alpha}|\widehat{\boldsymbol{\theta}}^{(k)}) \quad = \quad -\frac{1}{2}\log|\mathbf{D}| - \frac{1}{2}\mathrm{tr}\left(\widehat{u\mathbf{b}_i^2}^{(k)}\mathbf{D}^{-1}\right),$$

with $\widehat{a}_i^{(k)} = \mathrm{tr}\left(\widehat{u\mathbf{y}_i^2}^{(k)} - 2\widehat{u\mathbf{y}\mathbf{b}_i}^{(k)}\mathbf{Z}_i^\top + \widehat{u\mathbf{b}_i^2}^{(k)}\mathbf{Z}_i^\top\mathbf{Z}_i\right)$; $\widehat{u\mathbf{b}_i^2}^{(k)} = E\{u_i\mathbf{b}_i\mathbf{b}_i^\top|\mathbf{Q}_i, \mathbf{C}_i, \widehat{\boldsymbol{\theta}}^{(k)}\} = \widehat{\sigma^2}^{(k)}\widehat{\boldsymbol{\Lambda}}_i^{(k)} + \widehat{\boldsymbol{\varphi}}_i^{(k)}(\widehat{u\mathbf{y}_i^2}^{(k)} - \widehat{u\mathbf{y}}_i^{(k)}\widehat{\boldsymbol{\beta}}^{(k)\top}\mathbf{X}_i^\top - \mathbf{X}_i\widehat{\boldsymbol{\beta}}^{(k)}\widehat{u\mathbf{y}}_i^{(k)\top} + \widehat{u}_i^{(k)}\mathbf{X}_i\widehat{\boldsymbol{\beta}}^{(k)}\widehat{\boldsymbol{\beta}}^{(k)\top}\mathbf{X}_i^\top)\widehat{\boldsymbol{\varphi}}_i^\top$; $\widehat{u\mathbf{b}}_i^{(k)} = E\{u_i\mathbf{b}_i|\mathbf{Q}_i, \mathbf{C}_i, \widehat{\boldsymbol{\theta}}^{(k)}\} = \widehat{\boldsymbol{\varphi}}_i^{(k)}(\widehat{u\mathbf{y}}_i^{(k)} - \widehat{u}_i^{(k)}\mathbf{X}_i\widehat{\boldsymbol{\beta}}^{(k)})$; $\widehat{u\mathbf{y}\mathbf{b}_i}^{(k)} = E\{u_i\mathbf{y}_i\mathbf{b}_i^\top|\mathbf{Q}_i, \mathbf{C}_i, \widehat{\boldsymbol{\theta}}^{(k)}\} = (\widehat{u\mathbf{y}_i^2}^{(k)} - \widehat{u\mathbf{y}}_i^{(k)}\widehat{\boldsymbol{\beta}}^{(k)\top}\mathbf{X}_i^\top)\widehat{\boldsymbol{\varphi}}_i^\top$, where $\widehat{\boldsymbol{\Lambda}}_i^{(k)} = (\widehat{\sigma^2}^{(k)}\widehat{\mathbf{D}}^{-1(k)} + \mathbf{Z}_i^\top\mathbf{Z}_i)^{-1}$ and $\widehat{\boldsymbol{\varphi}}_i^{(k)} = \widehat{\boldsymbol{\Lambda}}_i^{(k)}\mathbf{Z}_i^\top$. Note that in this case we do not consider the computation of $E[h(u_i|\nu)|\mathbf{Q}, \mathbf{C}, \widehat{\boldsymbol{\theta}}^{(k)}]$, because ν is fixed.

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:

$$\widehat{\boldsymbol{\beta}}^{(k+1)} = \left(\sum_{i=1}^{n} \widehat{u}_i^{(k)} \mathbf{X}_i^\top \mathbf{X}_i\right)^{-1} \sum_{i=1}^{n} \mathbf{X}_i^\top \left(\widehat{u\mathbf{y}}_i^{(k)} - \mathbf{Z}_i \widehat{u\mathbf{b}}_i^{(k)}\right), \quad (18)$$

$$\widehat{\sigma^2}^{(k+1)} = \frac{1}{N}\sum_{i=1}^{n}\left[\widehat{a}_i^{(k)} - 2\widehat{\boldsymbol{\beta}}^{(k)\top}\mathbf{X}_i^\top(\widehat{u\mathbf{y}}_i^{(k)} - \mathbf{Z}_i\widehat{u\mathbf{b}}_i^{(k)}) + \widehat{u}_i^{(k)}\widehat{\boldsymbol{\beta}}^{(k)\top}\mathbf{X}_i^\top\mathbf{X}_i\widehat{\boldsymbol{\beta}}^{(k)}\right], \quad (19)$$

$$\widehat{\mathbf{D}}^{(k+1)} = \frac{1}{n}\sum_{i=1}^{n}\widehat{u\mathbf{b}_i^2}^{(k)}, \quad (20)$$

where $N=\sum_{i=1}^{n} n_i$. This process is iterated until some distance involving two successive evaluations of the log-likelihood $\ell(\boldsymbol{\theta}|\mathbf{y})$ described in Section 3.2, 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. That is, convergence is declared when the improvement in log-likelihood falls below a certain preset limit. In practice, *pmvt*() shows small random variability, which leads to nonincreasing log-likelihood beyond a certain level. The variability due to *pmvt*() can be controlled using the *algorithm* = *GenzBretz*(value) argument.

From (18)-(20) it is clear that the E-step reduces only to the computation of $\widehat{u\mathbf{y}_i^2}$, $\widehat{u\mathbf{y}}_i$ and $\widehat{u}_i$. These expected values can be determined in closed form, using propositions 1-3, as follows.

1. If $\mathbf{y}_i = \mathbf{y}_i^c$, i.e, the individual i has only censored components. Then from Proposition 2, we have:

$$\begin{aligned}
\widehat{u\mathbf{y}_i^2} &= E\{u_i\mathbf{y}_i\mathbf{y}_i^\top|\mathbf{Q}_i,\mathbf{C}_i,\widehat{\boldsymbol{\theta}}\} = \frac{T_{n_i}(\mathbf{Q}_i|\widehat{\boldsymbol{\mu}}_i,\widehat{\boldsymbol{\Sigma}}_i^*,\nu+2)}{T_{n_i}(\mathbf{Q}_i|\widehat{\boldsymbol{\mu}}_i,\widehat{\boldsymbol{\Sigma}}_i,\nu)}E\{\mathbf{W}_i\mathbf{W}_i^\top\},\\
\widehat{u\mathbf{y}}_i &= E\{u_i\mathbf{y}_i|\mathbf{Q}_i,\mathbf{C}_i,\widehat{\boldsymbol{\theta}}\} = \frac{T_{n_i}(\mathbf{Q}_i|\widehat{\boldsymbol{\mu}}_i,\widehat{\boldsymbol{\Sigma}}_i^*,\nu+2)}{T_{n_i}(\mathbf{Q}_i|\widehat{\boldsymbol{\mu}}_i,\widehat{\boldsymbol{\Sigma}}_i,\nu)}E\{\mathbf{W}_i\},\\
\widehat{u}_i &= E\{u_i|\mathbf{Q}_i,\mathbf{C}_i,\widehat{\boldsymbol{\theta}}\} = \frac{T_{n_i}(\mathbf{Q}_i|\widehat{\boldsymbol{\mu}}_i,\widehat{\boldsymbol{\Sigma}}_i^*,\nu+2)}{T_{n_i}(\mathbf{Q}_i|\widehat{\boldsymbol{\mu}}_i,\widehat{\boldsymbol{\Sigma}}_i,\nu)},
\end{aligned}$$

   where $\mathbf{W}_i \sim Tt_{n_i}(\widehat{\boldsymbol{\mu}}_i,\widehat{\boldsymbol{\Sigma}}_i^*,\nu+2;\mathbb{A}_i)$, $\widehat{\boldsymbol{\mu}}_i = \mathbf{X}_i\widehat{\boldsymbol{\beta}}$, $\widehat{\boldsymbol{\Sigma}}_i^* = \dfrac{\nu}{\nu+2}\widehat{\boldsymbol{\Sigma}}_i$, $\widehat{\boldsymbol{\Sigma}}_i = \widehat{\sigma^2}\mathbf{I}_{ni} + \mathbf{Z}_i\widehat{\mathbf{D}}\mathbf{Z}_i^\top$ and $\mathbb{A}_i = \{\mathbf{W}_i = (w_1,\ldots,w_{n_i})^\top | w_1 \le Q_{i1},\ldots,w_{n_i}\le Q_{in_i}\}$.

2. If $\mathbf{y}_i = \mathbf{y}_i^o$, i.e, the individual i has non censored components. Then,

$$\widehat{u\mathbf{y}_i^2} = \frac{\nu+n_i}{\nu+Q(\mathbf{y}_i)}\mathbf{y}_i\mathbf{y}_i^\top,\ \widehat{u\mathbf{y}}_i = \frac{\nu+n_i}{\nu+Q(\mathbf{y}_i)}\mathbf{y}_i,\ \widehat{u}_i = \frac{\nu+n_i}{\nu+Q(\mathbf{y}_i)},$$

   where $Q(\mathbf{y}_i) = (\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})^\top\boldsymbol{\Sigma}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})$, and finally

3. If $\mathbf{y}_i = (\mathbf{y}_i^{c\top}, \mathbf{y}_i^{o\top})^\top$, i.e., for individual i, we observed censored and uncensored components. Then from Proposition 3 and by the fact that $\{\mathbf{y}_i|\mathbf{Q}_i, \mathbf{C}_i\} = \{\mathbf{y}_i|\mathbf{Q}_i, \mathbf{C}_i, \mathbf{y}_i^o\} = \{\mathbf{y}_i^c|\mathbf{Q}_i, \mathbf{C}_i, \mathbf{y}_i^o\}$, we have

$$\begin{aligned}
\widehat{u\mathbf{y}_i^2} &= E\{u_i\mathbf{y}_i\mathbf{y}_i^\top|\mathbf{y}_i^o, \mathbf{Q}_i, \mathbf{C}_i, \widehat{\boldsymbol{\theta}}\} = \begin{pmatrix} \mathbf{y}_i^o\mathbf{y}_i^{o\top}\widehat{u}_i & \widehat{u}_i\mathbf{y}_i^o\widehat{\mathbf{w}}_i^{c\top} \\ \widehat{u}_i\widehat{\mathbf{w}}_i^c\mathbf{y}_i^{o\top} & \widehat{u}_i\widehat{\mathbf{w}^2}_i^c \end{pmatrix}, \\
\widehat{u\mathbf{y}}_i &= E\{u_i\mathbf{y}_i|\mathbf{y}_i^o, \mathbf{Q}_i, \mathbf{C}_i, \widehat{\boldsymbol{\theta}}\} = vec(y_i^o\widehat{u}_i, \widehat{\mathbf{w}}_i^c), \\
\widehat{u}_i &= E\{u_i|\mathbf{y}_i^o, \mathbf{Q}_i, \mathbf{C}_i, \widehat{\boldsymbol{\theta}}\} = \left(\frac{n_i^o + \nu}{\nu + Q(\mathbf{y}_i^o)}\right) \frac{T_p(\mathbf{Q}_i|\boldsymbol{\mu}_i^{co}, \widetilde{\mathbf{S}}^{co}, \nu + n_i^o + 2)}{T_p(\mathbf{Q}_i|\boldsymbol{\mu}_i^{co}, \mathbf{S}^{co}, \nu + n_i^o)},
\end{aligned}$$

where $\widetilde{\mathbf{S}}^{co} = \left(\dfrac{\nu + Q(\mathbf{y}_i^o)}{\nu + 2 + n_i^o}\right)\boldsymbol{\Sigma}_i^{cc.o}$, $\widehat{\mathbf{w}}_i^c = E\{\mathbf{W}_i\}$ and $\widehat{\mathbf{w}^2}_i^c = E\{\mathbf{W}_i\mathbf{W}_i^\top\}$, with $\mathbf{W}_i \sim Tt_{n_i^c}(\boldsymbol{\mu}_i^{co}, \widetilde{\mathbf{S}}^{co}, \nu + n_i^o + 2; \mathbb{A}_i^c)$ and $\boldsymbol{\Sigma}_i^{cc.o}$, $\boldsymbol{\mu}_i^{co}$ and $\mathbf{S}^{co}$ are as in (12)-(13).

3.4. Estimation of random effects and the expected information matrix

In this subsection we consider an empirical Bayes inference for the random effects, that is, the minimum mean squared error (MSE) predictor of $\mathbf{b}_i$, that is useful for evaluating subject-specific quantities such as individual intercepts and slopes. Thus, if values of parameter vector $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \sigma^2, \boldsymbol{\alpha}^\top)^\top$ and ν were known, the conditional mean of $\mathbf{b}_i$ given $\mathbf{C}_i$, $\mathbf{Q}_i$ is

$$\begin{aligned}
\widehat{\mathbf{b}}_i(\boldsymbol{\theta}) &= E\{\mathbf{b}_i|\mathbf{Q}_i, \mathbf{C}_i\} = E\{E\{E\{\mathbf{b}_i|u_i\}|\mathbf{y}_i, u_i\}|\mathbf{Q}_i, \mathbf{C}_i\} \\
&= E\{\boldsymbol{\Lambda}_i\mathbf{Z}_i^\top(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})|\mathbf{Q}_i, \mathbf{C}_i\} = \boldsymbol{\Lambda}_i\mathbf{Z}_i^\top(\widehat{\mathbf{y}}_i - \mathbf{X}_i\boldsymbol{\beta}), \qquad (21)
\end{aligned}$$

where $\boldsymbol{\Lambda}_i$ is defined in Section 3.3 and $\widehat{\mathbf{y}}_i = E\{\mathbf{y}_i|\mathbf{Q}_i, \mathbf{C}_i\}$ is the fist moment of the truncated multivariate-t distribution $(Tt_{n_i}(\mathbf{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma}_i, \nu; \mathbb{A}_i))$. In practice, the empirical Bayes estimators of $\mathbf{b}_i$, $\widehat{\mathbf{b}}_i$, can be obtained by substituting the ML estimate $\widehat{\boldsymbol{\theta}}$ into (21), which leads to $\widehat{\mathbf{b}}_i = \widehat{\mathbf{b}}_i(\widehat{\boldsymbol{\theta}})$. The conditional covariance matrix of $\mathbf{b}_i$ given $\mathbf{C}_i$, $\mathbf{Q}_i$ is

$$\begin{aligned}
Var\{\mathbf{b}_i|\mathbf{Q}_i, \mathbf{C}_i\} &= E\{\mathbf{b}_i\mathbf{b}_i^\top|\mathbf{Q}_i, \mathbf{C}_i\} - \widehat{\mathbf{b}}_i(\boldsymbol{\theta})\widehat{\mathbf{b}}_i(\boldsymbol{\theta})^\top \\
&= \frac{\nu + n_i}{\nu + n_i - 2}E\left\{(\frac{\nu + n_i}{\nu + Q(\mathbf{y}_i)})^{-1}|\mathbf{Q}_i, \mathbf{C}_i\right\}\boldsymbol{\Lambda}_i\sigma^2 + \boldsymbol{\Lambda}_i\mathbf{Z}_i^\top(\widehat{\mathbf{y}_i^2} - \widehat{\mathbf{y}}_i\widehat{\mathbf{y}}_i^\top)\mathbf{Z}_i\boldsymbol{\Lambda}_i,
\end{aligned}$$

where $\widehat{\mathbf{y}_i^2} = E\{\mathbf{y}_i\mathbf{y}_i^\top|\mathbf{Q}_i, \mathbf{C}_i\}$ is the second moment of the truncated multivariate-t distribution $(Tt_{n_i}(\mathbf{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma}_i, \nu; \mathbb{A}_i))$. These expected values can be easily accomplished from steps [1]-[3] given above as a by-product of our proposed ECM algorithm (E-step).

Louis (1982) derives a result that can be used to adjust the variances of the estimated fixed effects for the information lost due to censoring. Using this method, from the results given in Appendix B in Lange et al. (1989), an asymptotic approximation for the variances of the fixed effects is given by (see Appendix B):

$$\mathbf{J}_{\boldsymbol{\beta}\boldsymbol{\beta}} = Var(\widehat{\boldsymbol{\beta}}) = \left(\sum_{i=1}^{n} \frac{\nu + n_i}{\nu + n_i + 2} \mathbf{X}_i^\top \boldsymbol{\Sigma}_i^{-1} \mathbf{X}_i - \sum_{i=1}^{n} \mathbf{X}_i^\top \boldsymbol{\Sigma}_i^{-1} \mathbf{B}_i \boldsymbol{\Sigma}_i^{-1} \mathbf{X}_i \right)^{-1}, \tag{22}$$

where $\mathbf{B}_i = Var\left\{\frac{\nu + n_i}{\nu + Q(\mathbf{y}_i)}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})|\mathbf{Q}_i, \mathbf{C}_i\right\}$, with $\mathbf{y}_i \sim Tt_{n_i}(\mathbf{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma}_i, \nu; \mathbb{A}_i)$. Asymptotic confidence intervals and hypothesis tests for the fixed effects are obtained assuming that the MLE $\boldsymbol{\beta}$ has approximately a $N_p(\boldsymbol{\beta}, \mathbf{J}^{-1}_{\boldsymbol{\beta}\boldsymbol{\beta}})$ distribution. In practice, $\mathbf{J}_{\boldsymbol{\beta}\boldsymbol{\beta}}$ usually unknown and has to be replaced by its MLE $\widehat{\mathbf{J}}_{\boldsymbol{\beta}\boldsymbol{\beta}}$.

3.5. Model choice

A variety of information criteria exist to properly determine the best choice among a set of competing models. To identify the best selected model support by the data, we adopt the AIC and the BIC, which are the two most commonly used model selection tools and are of the form

$$-2\ell(\widehat{\boldsymbol{\theta}}|\mathbf{y}) + C(n)p,$$

where p is the number of parameters in the model and $C(n) = 2$ for the AIC and $C(n) = log(n)$ for the BIC. Both criteria can be applied to non-nested and to nested models, but not always lead to the same choice. Basically, there is no clear consensus regarding which criterion is better to use. A combined use of AIC and BIC would be of help to screening reasonable candidate models.

A formal test concerning the appropriateness of using the normal model $H_0 : \nu^{-1} = 0$ versus t model $H_1 : \nu^{-1} > 0$ is nontrivial since the null hypothesis is on the boundary of the parameter space. For testing parameters under non-standard settings, Self and Liang (1987) have shown the limiting distribution of the likelihood ratio test (LR) statistic will follow a mixture of chi-square distributions. Referring to Case 5 of Self and Liang (1987), the LR statistic under $H_0 : \nu^{-1} = 0$ is an equally weighted mixture of χ^2_0 and χ^2_1 distributions, where χ^2_0 denotes a degenerate distribution with all of its mass or probability at zero. In this case, the critical values are 1.65, 2.71 and 5.41 at the 10%, 5% and 1% significance levels, respectively.

4. More general linear mixed effects models

4.1. Random effects with different variance matrices

Pinheiro and Bates (2000) proposes different structures for $\mathbf{D}$ despite the unstructured matrix. Following Pinheiro and Bates (2000) and Vaida and Liu (2009) we present how to implement a variety of structures for $\mathbf{D}$. The new M-step are as follow:

(a) Unstructured $\mathbf{D}$. $\widehat{\mathbf{D}} = \Upsilon = \frac{1}{n}\sum_{i=1}^{n} \widehat{u\mathbf{b}_i^2}$, as (20).
(b) Diagonal $\mathbf{D}$. $\widehat{\mathbf{D}} = \mathrm{diag}(\frac{1}{n}\sum_{i=1}^{n} \widehat{u\mathbf{b}_{ij}^2}, j = 1, \ldots, q)$, where $\widehat{u\mathbf{b}_{ij}^2} = E(u_i b_{ij}^2 | \mathbf{Q}_i, \mathbf{C}_i, \widehat{\boldsymbol{\theta}})$.
(c) Block-diagonal $\mathbf{D}$. Define $\mathbf{D} = diag(\mathbf{D_1} \ldots, \mathbf{D_M}$, and let the corresponding submatrices of Υ be $\Upsilon_1, \ldots, \Upsilon_M$. Then $\widehat{\mathbf{D}} = \mathrm{diag}(\Upsilon_1, \ldots, \Upsilon_M)$
(d) Multiple of identity. Define $\mathbf{D} = \tau^2\mathbf{I}$. Then $\hat{\tau}_2 = \frac{1}{nq}\sum_{i=1}^{n}\sum_{j=1}^{q} \widehat{u\mathbf{b}_{ij}^2}$.
(e) Compound symmetry. Let $\mathbf{D} = \tau^2\mathbf{I} + \rho\mathbf{J}$, where $\mathbf{J}$ is a matrix of ones. Then $\mathbf{D} = \frac{\mathrm{tr}(\Upsilon)}{q}\mathbf{I} + \frac{sum(\Upsilon) - \mathrm{tr}(\Upsilon)}{q(q-1)}$, where $sum(\Upsilon) = \sum_{i,j}\Upsilon_{ij}$.

The proofs are ommited and are analogous to the ones presented in Vaida and Liu (2009) (Appendix A1).

4.2. Heteroscedastics Error

A more general linear mixed effects model is represented by

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i, \tag{23}$$

with the assumption that

$$\begin{pmatrix} \mathbf{b}_i \\ \boldsymbol{\epsilon}_i \end{pmatrix} \overset{ind.}{\sim} t_{n_i+q}\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \mathbf{D} & \mathbf{0} \\ \mathbf{0} & \sigma^2\boldsymbol{\Omega}_{n_i} \end{pmatrix}, \nu\right), i = 1, \ldots, n, \tag{24}$$

where $\boldsymbol{\Omega}_{n_i}$, for simplicity of notation $\boldsymbol{\Omega}_i$, are positive definite matrices parametrized by α_Ω, such that, $\boldsymbol{\Omega}_i = \boldsymbol{\Omega}(\alpha_\Omega)_i$. Therefore, when $\boldsymbol{\Omega}_i = \mathbf{I}_{n_i}$ we recover the model presented in Section 3.1. From (24) it follows that

$$\mathbf{b}_i \overset{iid}{\sim} t_q(\mathbf{0}, \mathbf{D}, \nu) \quad \text{and} \quad \boldsymbol{\epsilon}_i \overset{iid}{\sim} t_{n_i}(\mathbf{0}, \sigma^2\boldsymbol{\Omega}_i, \nu), \quad i = 1, \ldots, n. \tag{25}$$

Define $\boldsymbol{\Omega}_i^{-1/2}$ the inverse of $\boldsymbol{\Omega}_i^{1/2}$ which is the square root of $\boldsymbol{\Omega}_i$, thus $\boldsymbol{\Omega} = (\boldsymbol{\Omega}_i^{1/2})^\top\boldsymbol{\Omega}_i^{1/2}$. Using $\boldsymbol{\Omega}_i^{-1/2}$ let

$$\mathbf{y}_\mathbf{i}^\star = (\boldsymbol{\Omega}_\mathbf{i}^{\mathbf{1/2}})^\top\mathbf{y}_\mathbf{i}, \quad \boldsymbol{\epsilon}_\mathbf{i}^\star = (\boldsymbol{\Omega}_\mathbf{i}^{\mathbf{1/2}})^\top\boldsymbol{\epsilon}_\mathbf{i},$$
$$\mathbf{X}_i^\star = (\boldsymbol{\Omega}_i^{1/2})^\top\mathbf{X}_i, \quad \mathbf{Z}_i^\star = (\boldsymbol{\Omega}_i^{1/2})^\top\mathbf{Z}_i.$$

Then, given the parameters α_Ω we have that

$$\mathbf{y}_i^\star = \mathbf{X}_i^\star\boldsymbol{\beta} + \mathbf{Z}_i^\star\mathbf{b}_i + \boldsymbol{\epsilon}_i^\star, \quad \mathbf{b}_i \overset{iid}{\sim} t_q(\mathbf{0}, \mathbf{D}, \nu) \quad \text{and} \quad \boldsymbol{\epsilon}_i^\star \overset{iid}{\sim} t_{n_i}(\mathbf{0}, \sigma^2\mathbf{I}_{n_i}, \nu), i = 1, \ldots, n,$$

which is the linear mixed effect model presented in (7) and (8).

Following Vaida and Liu (2009), to estimate α_Ω we parametrize $\boldsymbol{\Omega}_i = \mathbf{V}_i\mathbf{K}_i\mathbf{V}_i$, where $\mathbf{V}_i$ is diag($\sqrt{\text{var}(\epsilon_1)/\sigma^2}, \ldots, \sqrt{\text{var}(\epsilon_{n_i})/\sigma^2}$), $\mathbf{K}_i = Corr(\epsilon_i)$ and assume that $\alpha_\Omega = vec(\alpha_v, \alpha_k)$. This assumption is equivalent to assume that the parameters in $\mathbf{V}_i$ and $\mathbf{K}_i$ are independent. Thus, $\mathbf{V}_i = \mathbf{V}(\alpha_v)_i$ and $\mathbf{K}_i = \mathbf{K}(\alpha_k)_i$. Such assumption allows α_v and α_k to be estimated separately depending on the model asuumed for the variance function and within-subject correlation structure.

To exemplify how it is done, suppose that $\boldsymbol{\Omega}_i$ has a first order autoregressive strucuture (AR(1)), i. e.,

$$\begin{bmatrix} 1, \rho^{|2-1|}, \ldots, \rho^{|q-1|} \\ \rho^{|1-2|}, 1, \ldots, \rho^{|q-2|} \\ \vdots, \ldots, \ddots, \vdots \\ \rho^{|1-q|}, \rho^{|2-q|}, \ldots, 1 \end{bmatrix}.$$

The parameter ρ is updated solving

$$\frac{1}{2}\sum_{i=1}^{n} \text{tr}\{[\boldsymbol{\Omega}_i^{-1} - \sigma^{-2}\boldsymbol{\Omega}_i^{-1}E(u_i\boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i^\top|\mathbf{Q}_i, \mathbf{C}_i, \widehat{\boldsymbol{\theta}})\boldsymbol{\Omega}_i^{-1}]\dot{\boldsymbol{\Omega}}_{\mathbf{i}}\} = 0,$$

where $\dot{\boldsymbol{\Omega}}_{\mathbf{i}} = \partial\boldsymbol{\Omega}_i/\partial\rho$ and

$$E(u_i\boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i^\top|\mathbf{Q}_i, \mathbf{C}_i, \widehat{\boldsymbol{\theta}}) = E(u_i(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{Z}_i\mathbf{b}_i)(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{Z}_i\mathbf{b}_i)^\top|\mathbf{Q}_i, \mathbf{C}_i, \widehat{\boldsymbol{\theta}}),$$

which can be estimated following the steps to solve equation (17) (for more details see appendix 8). Vaida and Liu (2009) (Subsection 3.2) presents others examples of variance functions that can be extended for our formulation.

5. The nonlinear case

Extending the notation of the previous section and ignoring censoring, we first propose the following general mixed-effects model in which the random terms are assumed to follow a multivariate-t distribution (t-NLME). Let $\mathbf{y}_i = (y_{i1}, \ldots, y_{in_i})^\top$ denote the (continuous) response vector for subject i and $\eta = (\eta(\mathbf{X}_{i1}, \boldsymbol{\phi}_i), \ldots, \eta(\mathbf{X}_{in_i}, \boldsymbol{\phi}_i))^\top$ be a nonlinear vectorvalued differentiable function of the individuals random parameter $\boldsymbol{\phi}_i$ and a vector of covariates $\mathbf{X}_i$. The t-NLME can then be expressed as:

$$\mathbf{y}_i = \eta(\boldsymbol{\phi}_i, \mathbf{X}_i) + \boldsymbol{\epsilon}_i, \quad \boldsymbol{\phi}_i = \mathbf{A}_i\boldsymbol{\beta} + \mathbf{B}_i\mathbf{b}_i, \tag{26}$$

where the joint distribution of ($\mathbf{b}_i$, $\boldsymbol{\epsilon}_i$) is as in (8), $\mathbf{A}_i$ and $\mathbf{B}_i$ are known design matrices of dimensions $r \times p$ and $r \times q$ respectively, possibly depending on some covariable values, $\boldsymbol{\beta}$ is the $(p \times 1)$ vector of fixed effects, $\mathbf{b}_i$ is the $(q \times 1)$ vector of random effects. Thus, from the properties of the multivariate-t distribution, we have that marginally,

$$\boldsymbol{\phi}_i \overset{ind}{\sim} t_r(\mathbf{A}_i\boldsymbol{\beta}, \mathbf{B}_i\mathbf{D}\mathbf{B}_i^{\top}, \nu) \text{ and } \boldsymbol{\epsilon}_i \overset{ind.}{\sim} t_{n_i}(\mathbf{0}, \sigma^2\mathbf{I}_{n_i}, \nu), \tag{27}$$

and as in the linear case, they are uncorrelated because $\text{Cov}(\boldsymbol{\phi}_i, \boldsymbol{\epsilon}_i) = \mathbf{0}$. For NI-NLME with non censoring responses, the marginal distribution is given by

$$f(\mathbf{y}|\boldsymbol{\theta}) = \prod_{i=1}^{n}\int_0^{\infty}\int_{\mathbb{R}^q} \phi_{n_i}(\mathbf{y}_i; \eta(\boldsymbol{\phi}_i, \mathbf{X}_i), u_i^{-1}\sigma^2\mathbf{I}_{n_i})\phi_q(\boldsymbol{\phi}_i; \mathbf{A}_i\boldsymbol{\beta}, u_i^{-1}\mathbf{B}_i\mathbf{D}\mathbf{B}_i^{\top}) \times G(u_i|\nu/2, \nu/2)d\boldsymbol{\phi}_i du_i, \tag{28}$$

which generally does not have a closed form expression because the model function is not linear in the random effect. In the normal case, various approximations (viz. first-order Taylor series expansion of the model function around the conditional mode of $\mathbf{b}_i$, says $\widetilde{\mathbf{b}}_i$) have been proposed to achieve tractable numerical optimizations (Wu, 2010). Most algorithms for computing the approximate MLE $\widehat{\boldsymbol{\theta}}$ and empirical Bayes estimators (predictors) for the random effects $\widehat{\mathbf{b}}_i$ considers iterative maximization of the approximate log-likelihood functions $\ell(\boldsymbol{\theta}, \widetilde{\mathbf{b}}) = \sum_{i=1}^{n} \log f(\mathbf{y}_i|\boldsymbol{\theta}, \widetilde{\mathbf{b}}_i)$. Following Taylor series expansions, we have the following theorems. The first uses a point in a neighborhood of the conditional mode $\widetilde{\mathbf{b}}_i$ as the expansion point and it has been proven useful for implementation of model selection, in a Bayesian context (Lachos et al., 2011). The second, useful for the implementation of the EM algorithm, uses simultaneously neighborhood of $\mathbf{b}_i$ and $\boldsymbol{\beta}$ as expansions points, with the advantage that the likelihood is completely linearized (in $\mathbf{b}_i$ and $\boldsymbol{\beta}$). We call these LME approximations and can be considered as extensions of the result given in Lindstrom and Bates (1990) and Pinheiro and Bates (2000) for the Student-t case.

THEOREM 1. *Let $\widetilde{\mathbf{b}}_i$ be an expansion point in a neighborhood of $\mathbf{b}_i$, then under the t–NLME model as in (26)-(27), the marginal distribution of $\mathbf{y}_i$, can be approximated as* $\mathbf{y}_i \dot{\sim} t_{ni}(\eta(\mathbf{A}_i\boldsymbol{\beta} + \mathbf{B}_i\widetilde{\mathbf{b}}_i, \mathbf{X}_i) - \widetilde{\mathbf{H}}_i\widetilde{\mathbf{b}}_i, \widetilde{\mathbf{V}}_i, \nu)$, *where* $\widetilde{\mathbf{V}}_i = (\widetilde{\mathbf{H}}_i\mathbf{B}_i)\mathbf{D}(\widetilde{\mathbf{H}}_i\mathbf{B}_i)^{\top} + \sigma^2\mathbf{I}_{n_i}$, $\widetilde{\mathbf{H}}_i = \dfrac{\partial\eta(\mathbf{A}_i\boldsymbol{\beta} + \mathbf{B}_i\mathbf{b}_i, \mathbf{X}_i)}{\partial\mathbf{b}_i^{\top}}|_{\mathbf{b}_i=\widetilde{\mathbf{b}}_i}$ *and* $\dot{\sim}$ *denotes approximated in distribution.*

PROOF. See Lachos et al. (2011).

The next theorem allows the implementation of the EM algorithm.

THEOREM 2. *Let $\widetilde{\mathbf{b}}_i$ and $\widetilde{\boldsymbol{\beta}}$ be expansion points in a neighborhood of $\mathbf{b}_i$ and $\boldsymbol{\beta}$, respectively, then under the t-NLME model as (26)–(27), we have the following linearized model*

$$\widetilde{\mathbf{y}}_i = \widetilde{\mathbf{W}}_i\boldsymbol{\beta} + \widetilde{\mathbf{H}}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i, \;\; i = 1, \ldots, n, \tag{29}$$

where $\widetilde{\mathbf{y}}_i = \mathbf{y}_i - \widetilde{\eta}(\mathbf{A}_i\widetilde{\boldsymbol{\beta}} + \mathbf{B}_i\widetilde{\mathbf{b}}_i, \mathbf{X}_i)$, $\mathbf{b}_i \overset{ind}{\sim} t_q(0, \mathbf{D}, \nu)$ *and* $\boldsymbol{\epsilon}_i \overset{ind.}{\sim} t_{n_i}(\mathbf{0}, \sigma^2\mathbf{I}_{n_i}, \nu)$, $\widetilde{\mathbf{H}}_i = \dfrac{\partial\eta(\mathbf{A}_i\boldsymbol{\beta} + \mathbf{B}_i\mathbf{b}_i, \mathbf{X}_i)}{\partial\mathbf{b}_i^\top}|_{\mathbf{b}_i=\widetilde{\mathbf{b}}_i}$ *and* $\widetilde{\mathbf{W}}_i = \dfrac{\partial\eta(\mathbf{A}_i\boldsymbol{\beta} + \mathbf{B}_i\mathbf{b}_i, \mathbf{X}_i)}{\partial\boldsymbol{\beta}_i^\top}|_{\boldsymbol{\beta}_i=\widetilde{\boldsymbol{\beta}}_i}$ *and* $\widetilde{\eta}(\widetilde{\boldsymbol{\beta}}, \widetilde{\mathbf{b}}_i) = \eta(\mathbf{A}_i\widetilde{\boldsymbol{\beta}} + \mathbf{B}_i\widetilde{\mathbf{b}}_i, \mathbf{X}_i) - \widetilde{\mathbf{H}}_i\widetilde{\mathbf{b}}_i - \widetilde{\mathbf{W}}_i\widetilde{\boldsymbol{\beta}}$,

PROOF. Based on first-order Taylor expansion of the function η around $\widetilde{\mathbf{b}}_i$ and $\widetilde{\boldsymbol{\beta}}$, we have that

$$\eta(\mathbf{A}_i\boldsymbol{\beta} + \mathbf{B}_i\mathbf{b}_i, \mathbf{X}_i) \approx [\eta(\mathbf{A}_i\widetilde{\boldsymbol{\beta}} + \mathbf{B}_i\widetilde{\mathbf{b}}_i, \mathbf{X}_i) + \widetilde{\mathbf{H}}_i\mathbf{b}_i - \widetilde{\mathbf{H}}_i\widetilde{\mathbf{b}}_i + \widetilde{\mathbf{W}}_i\boldsymbol{\beta} - \widetilde{\mathbf{W}}_i\widetilde{\boldsymbol{\beta}}]$$

with $\widetilde{\mathbf{H}}_i = \dfrac{\partial\eta(\mathbf{A}_i\boldsymbol{\beta} + \mathbf{B}_i\mathbf{b}_i, \mathbf{X}_i)}{\partial\mathbf{b}_i^\top}|_{\mathbf{b}_i=\widetilde{\mathbf{b}}_i}$ and $\widetilde{\mathbf{W}}_i = \dfrac{\partial\eta(\mathbf{A}_i\boldsymbol{\beta} + \mathbf{B}_i\mathbf{b}_i, \mathbf{X}_i)}{\partial\boldsymbol{\beta}_i^\top}|_{\boldsymbol{\beta}_i=\widetilde{\boldsymbol{\beta}}_i}$. It follows that

$$\begin{aligned}
\boldsymbol{\epsilon}_i &= \mathbf{y}_i - \eta(\mathbf{A}_i\boldsymbol{\beta} + \mathbf{B}_i\mathbf{b}_i, \mathbf{X}_i) \approx \mathbf{y}_i - [\eta(\mathbf{A}_i\widetilde{\boldsymbol{\beta}} + \mathbf{B}_i\widetilde{\mathbf{b}}_i, \mathbf{X}_i) + \widetilde{\mathbf{H}}_i\mathbf{b}_i - \widetilde{\mathbf{H}}_i\widetilde{\mathbf{b}}_i + \widetilde{\mathbf{W}}_i\boldsymbol{\beta} - \widetilde{\mathbf{W}}_i\widetilde{\boldsymbol{\beta}}] \\
&= \mathbf{y}_i - [\widetilde{\eta}(\widetilde{\boldsymbol{\beta}}, \widetilde{\mathbf{b}}_i) + \widetilde{\mathbf{W}}_i\boldsymbol{\beta} + \widetilde{\mathbf{H}}_i\mathbf{b}_i] = \widetilde{\mathbf{y}}_i - [\widetilde{\mathbf{W}}_i\boldsymbol{\beta} + \widetilde{\mathbf{H}}_i\mathbf{b}_i],
\end{aligned}$$

which concludes the proof.

The empirical Bayes estimates of the random effects $\widetilde{\mathbf{b}}$, given in (21), can be used iteratively in the linearization procedure from Theorem 2. Note that the distribution of $\mathbf{b}_i|\mathbf{y}_i$ is approximately symmetric (Student-t), and thus $\widetilde{\mathbf{b}}_i$ is the mode of the distribution at each step. As commented by Vaida and Liu (2009), the linearization (L) procedure to obtain the approximate MLE of $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \sigma^2, \boldsymbol{\alpha}^\top)^\top$ consists to iteratively solving the LME model (L-step) in (29). For censored response the linearized model (29) is an LME with censored data, with same structure as (7)-(8), which is then solved as indicated in the previous section. The model matrices in (29) depends on the current parameter value, and needs to be recalculated at each iteration. The algorithm iterates to convergence between L-, E-, and CM-steps.

6. Application

We illustrate the proposed methods with the analysis of two HIV datasets previously analyzed using N-LMEC models.

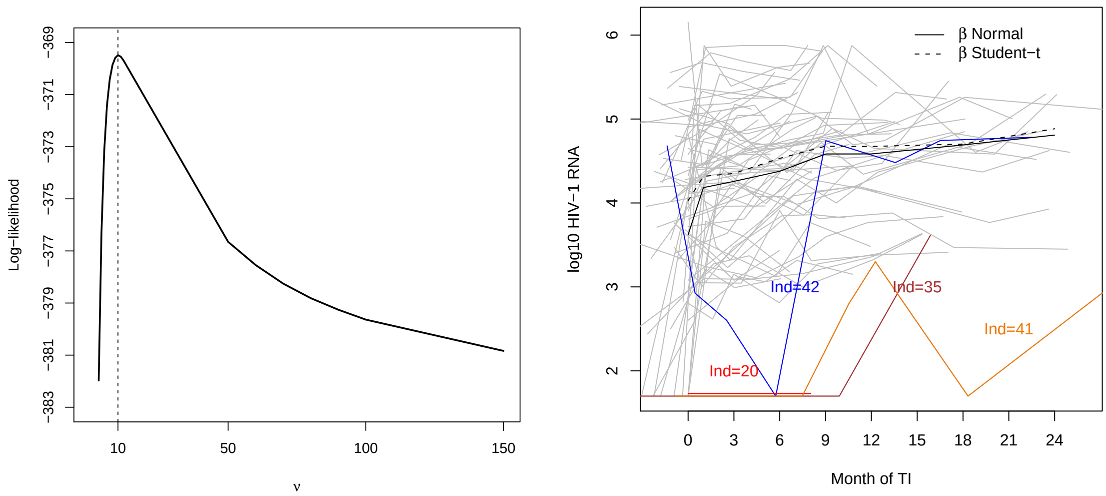


Fig. 1. UTI data. (Left panel) plot of the profile log-likelihood of the degrees of freedom ν. (Right panel) Individual profiles and overall mean (in $\log_{10}$ scale) using the Normal and t distributions for HIV viral load at different follow-up times. The trajectories for the influential individuals are numbered.

6.1. UTI Data

The first application is a study of 72 perinatally HIV-infected children (Saitoh et al., 2008). The data set is available in the R package *lmec*. Primarily due to treatment fatigue, unstructured treatment interruptions (UTI) is common in this population. Suboptimal adherence can lead to ARV resistance and diminished treatment options in the future. The subjects in the study had taken ARV therapy for at least 6 months before UTI, and the medication was discontinued for more than 3 months. Out of 362 observations, 26 (7%) observations were below the detection limits (50 or 400 copies/mL) and considered left-censored at these values. The individual profiles of viral load at different followup times after UTI is presented in Figure 1 (right panel). We consider a profile LME model with random intercepts b_i as $y_{ij} = b_i + \beta_j + \epsilon_{ij}$, where y_{ij} is the $\log_{10}$ HIV RNA for subject i at time t_j,$t_1 = 0$, $t_2 = 1$, $t_3 = 3$, $t_4 = 6$, $t_5 = 9$, $t_6 = 12$, $t_7 = 18$, $t_8 = 24$. Vaida and Liu (2009) analyzed the same data set by fitting a similar N-LMEC via the EM algorithm, but from Figure 1 given in Lachos et al. (2011) it is clear that inference based on normality assumptions are questionable (presence of thick tails). Thus, we revisit the UTI data with the aim of providing robust inferences, from a frequentist perspective, by using the Student-t distribution. The ML estimates were obtained using the ECM algorithm described in Sec-

Table 1. ML estimates under normal and Student-t models fitted to the UTI data. SE are the corresponding standard errors.

	N-LMEC		T-LMEC	
Parameter	estimate	SE	estimate	SE
β_1	3.6038	0.1253	3.6182	0.1238
β_2	4.1664	0.1285	4.2532	0.1311
β_3	4.2413	0.1304	4.3137	0.1332
β_4	4.3604	0.1307	4.4580	0.1338
β_5	4.5662	0.1398	4.6229	0.1435
β_6	4.5692	0.1485	4.6112	0.1532
β_7	4.6773	0.1646	4.6978	0.1709
β_8	4.7935	0.2018	4.7874	0.2111
σ^2	0.3414		0.3503	
α	0.7653		0.6662	
ν	-	-	10	-
AIC	844.1172		759.0148	
BIC	883.0337		797.9312	

tion 3. Starting values were obtained by using the library *lmec*.

For the Student-t model, we assumed that the degree of freedom ν is know and by using the AIC criterion we found $\nu = 10$ (see left panel in Figure 1). It is a first indication that the normal model is inadequate. Table 1 presents the ML estimate of $\boldsymbol{\theta}$ and the corresponding standard errors of the fixed effects. Comparing these values we notice a similarity between the estimates under normal and Student-t models. Additionally, the inferences for the variance components are similar for the two models, but are not comparable since they are on different scales. According to the AIC or BIC values, given at the bottom of Table 1, we notice also that the t-LMEC model perform better than the N-LMEC model. For the LR statistics described in Subsection 3.5, we have that the maximum log-likelihood for the N-LMEC model is -412.059 and for the t-LMEC model is -369.507, corresponding to a likelihood ratio statistics of $LR = 42.552$. Here the LR statistic follows a equally weighted mixture of χ^2_0 and χ^2_1 distributions. Therefore, the resulting p-value 3.441×10^{-11} guarantees the appropriateness of the use of the multivariate-t distribution. With missing-at-random assumption as in Vaida and Liu (2009), our dropout (censored) model does not bias the inference regarding the mean of β_j. For both models the mean viral load $E(y_{ij}) = \beta_j$ increases gradually throughout 24 months for the two models. For the best model (t-LMEC), it increases from 3.62 at the time of UTI to 4.79 at 24 months. The estimates of the between-subject (α) and within-subject (σ^2) scale parameters (in log10 scale) are 0.6662 and 0.3503, respectively.

To determine possible influential observations, we use the Mahalanobis distance $d_i^2(\boldsymbol{\theta}) =$

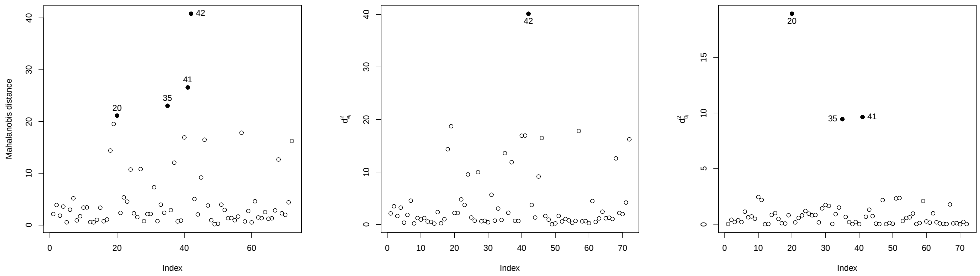


Fig. 2. UTI data. (a) Mahalanobis distance, (b) Estimated $d^2_{\mathbf{e}_i}$ (error) and (c) Estimated $d^2_{\mathbf{b}_i}$ (R.E.), for the N-LMEC model.

$(\widehat{\mathbf{y}}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}})^\top \boldsymbol{\Sigma}_i^{-1}(\widehat{\mathbf{y}}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}})$, $i = 1, \ldots, 72$. As in Pinheiro et al. (2001), replacing $\boldsymbol{\theta}$ and $\mathbf{b}_i$ with their current estimates, we obtain the following decomposition for the Mahalanobis distance:

$$\begin{aligned} d_i^2(\widehat{\boldsymbol{\theta}}) &= (\widehat{\mathbf{y}}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}})^\top(\widehat{\sigma^2}\mathbf{I}_{ni} + \mathbf{Z}_i^\top\widehat{\mathbf{D}}\mathbf{Z}_i)^{-1}(\widehat{\mathbf{y}}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}}) \\ &= -\frac{1}{\widehat{\sigma^2}}\widehat{\mathbf{e}_i}^\top\widehat{\mathbf{e}_i} + \widehat{\mathbf{b}_i}^\top\widehat{\mathbf{D}}\widehat{\mathbf{b}}_i, = \widehat{d^2_{\mathbf{e}_i}} + \widehat{d^2_{\mathbf{b}_i}} \end{aligned}$$

where $\widehat{\mathbf{e}_i} = \widehat{\mathbf{y}}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}} - \mathbf{Z}_i\widehat{\mathbf{b}_i}$ where $\widehat{\mathbf{b}_i}$ is as in (21). The estimated distances $d^2_{\mathbf{e}_i}$ (Error) and $d^2_{\mathbf{b}_i}$ (Random Effect-R.E.) provide a useful diagnostic statistics for identifying subjects with outlying observations (see, for example, Meza et al., 2011). Figure 2 presents these diagnostic statistics for N-LMEC model. Subjects #42 present large values of d_i^2 and $d^2_{\mathbf{e}_i}$, suggesting an outlying observation at the within-subject level ($\mathbf{e}$-outlier). Moreover, observations #20, #35 and #41 presents large value of $d^2_{\mathbf{b}_i}$, suggesting outlying observations at the between-subject level ($\mathbf{b}$-outlier). Under a Bayesian paradigm, these observations were also detected as influential in the work by Lachos et al. (2011).

It is well known that outlying observations may affect the estimation of the parameters under assumptions of normality. However, when we use the Student-t distribution, the EM algorithm allows to accommodate these discrepant observations attributing to them small weights in the estimation procedure. The estimated weights ($\widehat{u}_i$, $i = 1, \ldots, 72$) for the t-LMEC model are presented in Figure 3. We observe from this Figure that observations #20, #35, #41 and #42, indicated as outliers under the normal model, take the smaller values, confirming the robust aspects of the MLE against outlying observations under the t-LMEC model. The robustness of the t-LMEC is also observed in Figure 1 (right panel),

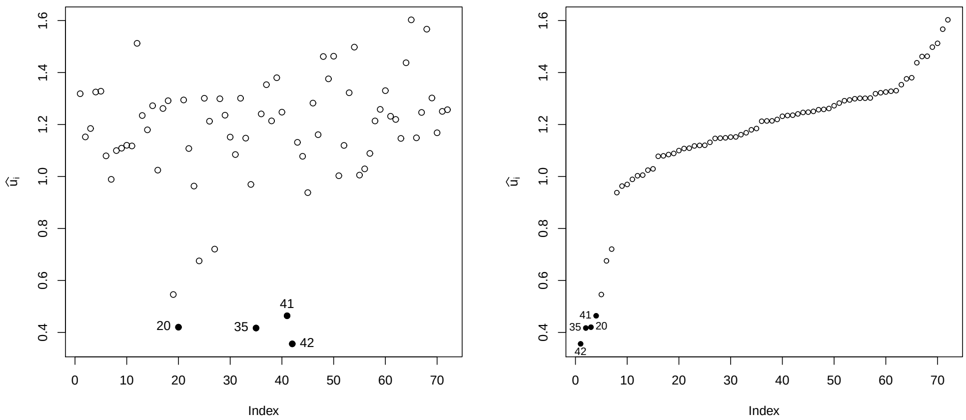


Fig. 3. UTI data. Estimated weight $\widehat{u}_i$ for the t-LMEC fit. The influential observations for the N-LMEC are numbered.

where the presence of these outliers might have underestimated the predicted mean curve for the N-LMEC model as compared to the t-LMEC model. In summary, we can see from this example that the robust aspects of the t-LME models (Pinheiro et al., 2001) against outlying observations are also extended to the case in which censoring components are present.

6.2. AIEDRP study

The second AIDS case study is from the AIEDRP program, a large multicenter observational study of subjects with acute and early HIV infection. We consider 320 untreated individuals with acute HIV infection; for more details on this dataset see Vaida and Liu (2009). Of the 830 recorded observations, 185 (22%) were above the limit of assay quantification, hence they were considered as right-censored. So, we consider a right-censored version and accommodate it within our NLME. Following Vaida and Liu (2009), we choose a five-parameter NLME model (inverted S-shaped curve) as follows:

$$y_{ij} = \alpha_{1i} + \frac{\alpha_2}{(1 + \exp((t_{ij} - \alpha_3)/\alpha_4))} + \alpha_{5i}(t_{ij} - 50) + \epsilon_{ij},$$

where y_{ij} is the $\log_{10}$ HIV RNA for subject i at time t_{ij}. The parameter α_{1i} and α_2 represents subject-specific (random) set points and decrease from the maximum HIV RNA. In the absence of treatment (following acute infection), the HIV RNA varies around a set-point

Table 2. ML estimates under normal and Student-t models fitted to the AIEDRP data. SE are the corresponding standard errors.

	N-LMEC		T-LMEC	
Parameter	estimate	SE	estimate	SE
β_1	1.60964	0.0147	1.61148	0.0133
β_2	0.14217	0.0949	0.16122	0.0849
β_3	3.52617	0.0237	3.52370	0.0208
β_4	1.05585	0.2677	0.98713	0.2458
β_5	-0.0035	0.0014	-0.0031	0.0013
σ^2	0.26521		0.20726	
α_{11}	0.01769		0.01611	
α_{12}	0.00016		0.00013	
α_{22}	0.00004		0.00004	
ν	-	-	10	-
AIC	1610.814		1581.416	
BIC	1700.521		1623.908	

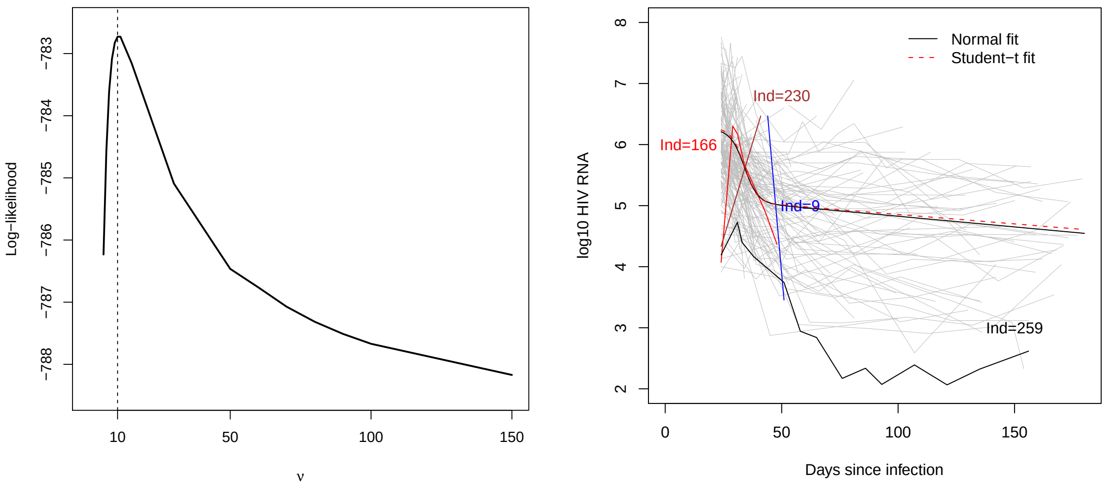


Fig. 4. AIEDRP data. (Left panel) plot of the profile log-likelihood of the degrees of freedom ν. (Right panel) Individual profiles and overall mean (in $\log_{10}$ scale) using the Normal and t distributions for HIV viral load at different follow-up times. The trajectories for the influential individuals are numbered.

which may differ among individuals, hence the set point is chosen to be subject-specific. The location parameter α_3 indicates the time point at which half of the change in HIV RNA is attained, α_4 is a scale parameter modeling the rate of decline and α_{5i} allows for increasing HIV RNA trajectory after day 50. To force the parameters to be positive, we reparameterize as follows: $\beta_{1i} = \log(\alpha_{1i}) = \beta_1 + b_{1i}$; $\beta_k = \log(\alpha_k), k = 2, 3, 4$ and $\alpha_{5i} = \beta_5 + b_{2i}$. Within a classical framework, we use the Student-t (t-NLMEC) with the ECM algorithm as described in Section 3. As in the previous application, the estimation of the parameters ν was chosen following the strategy proposed by Lange et al. (1989), which selects a small value for $\nu = 10$ (see left panel in Figure 4). This parameter act as tuning constant in robust estimation methods and in our case we see that this choice provide adequate protection against outliers. For the sake of model comparison, we also fit the N-NLMEC counterparts, which can be treated as the reduced t-NLMEC as ν tends to infinity.

Table 2 lists the ML estimates parameters for the N-LMEC model and the t-LMEC model, together with the corresponding standard errors of the fixed effects and the associated AIC and BIC values. From this table, we observe that the standard errors of the t-NLMEC are smaller, indicating that the Student-t model to produce more precise estimates. According to the AIC or BIC values, the t-NLMEC provided much improved model fits over the N-NLMEC. In fact, the maximum log-likelihood for the N-LMEC is -781.708 and for the t-LMEC model is -775.951, corresponding to the likelihood ratio statistics of 11.508 (p-value = 0.00035), this also reinforce the conclusion that the t-LMEC model fits the data significantly better than N-LMEC model.

To identify outlying observations, we compute the Mahalanobis distance $d_i^2(\widehat{\boldsymbol{\theta}})$, $i = 1, \ldots, 320$, the estimated distances $d^2_{\mathbf{e}_i}$ (Error) and $d^2_{\mathbf{b}_i}$ (Random Effect), were also computed for the normal case. Figure 5 presents these diagnostic statistics for the N-LMEC model. We can see from this figures that observations #9, #166, #230 and #259 appear as possibles outliers. The observations #9, #166 and #230 presents large value of $d^2_{\mathbf{e}_i}$, suggesting an $\mathbf{e}$-outlier. Moreover, observation #259 presents large value of $d^2_{\mathbf{b}_i}$, suggesting an $\mathbf{b}$-outlier. From figure 4 (right panel), the fitted viral load curve appears to be underestimated as compared to the t-NLMEC due to the presence of these outliers. This suggests that t-NLMEC, which downweights the influence of outliers, provides an appropriate way for achieving robust inference.

The robustness of the t-LMEC model can be assessed by considering the influence of a single outlying observation on the ML estimate of θ. In particular, we can assess how much the ML estimates of θ influences by a change of δ units in a single observation y_{ik}.

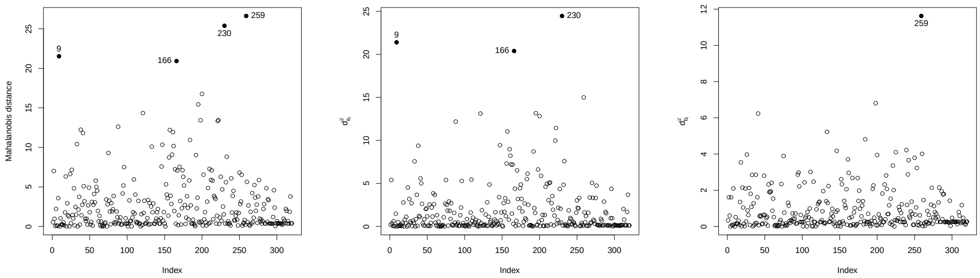


Fig. 5. AIEDRP data. (a) Mahalanobis distance, (b) Estimated $d^2_{\mathbf{e}_i}$ (error) and (c) Estimated $d^2_{\mathbf{b}_i}$ (R.E.). The influential observations are numbered.

We replace a single observation y_{ik} by $y_{ik}(\delta) = y_{ik} + \delta$, and record the relative change in the estimates $((\widehat{\theta}(\delta) - \widehat{\theta})/\widehat{\theta})$, where $\widehat{\theta}$ denotes the original estimate and $\widehat{\theta}(\delta)$ the estimate for the contaminated data. In this application we contaminated the first observation on subject 198 and varied δ between -10 and 10. In Figure 6 we present the results of the relatives changes of the estimates $\boldsymbol{\beta}$ and σ^2 for different values of δ, under the N-NLMEC and t-NLMEC models. As expected, the estimates from the t-NLMEC is less affected by variations of δ than the N-NLMEC.

7. Simulation studies

To study the performance of our proposed methodology we conduct a simulation study to illustrate the linear and nonlinear cases. The goal of this simulation study is to investigate the consequences on parameter inference when the normality assumption is inappropriate as well as to investigate whether the model comparison measures, AIC and BIC determines the best-fitting model to the simulated data.

7.1. The linear case

To study the linear regression, we consider the following linear mixed model:

$$y_{ij} = \beta_0 + \beta_1 t_{ij} + b_{0i} + b_{1i} t_{ij} + \epsilon_{ij}, \quad i = 1, \dots, 100, \quad j = 1, \dots, 6, \tag{30}$$

where $(b_{0i}, b_{1i}) \overset{\text{iid.}}{\sim} t_2(\mathbf{0}, \mathbf{D}, \nu)$, $\epsilon_{ij} \sim t(0, \sigma^2, \nu)$. We set $t_{ij} = (2, 4, 6, 8, 10, 24)$, $\boldsymbol{\beta}^\top = (\beta_0, \beta_1) = (-2.83, -0.18)$, $\mathbf{D} = \begin{bmatrix} 0.049 & 0.001 \\ 0.001 & 0.002 \end{bmatrix}$, $\sigma^2 = 0.15$ and $\nu = 4$.

We choose various settings of censoring proportions, 5%, $10\%, 20\%$ and 50%, to study

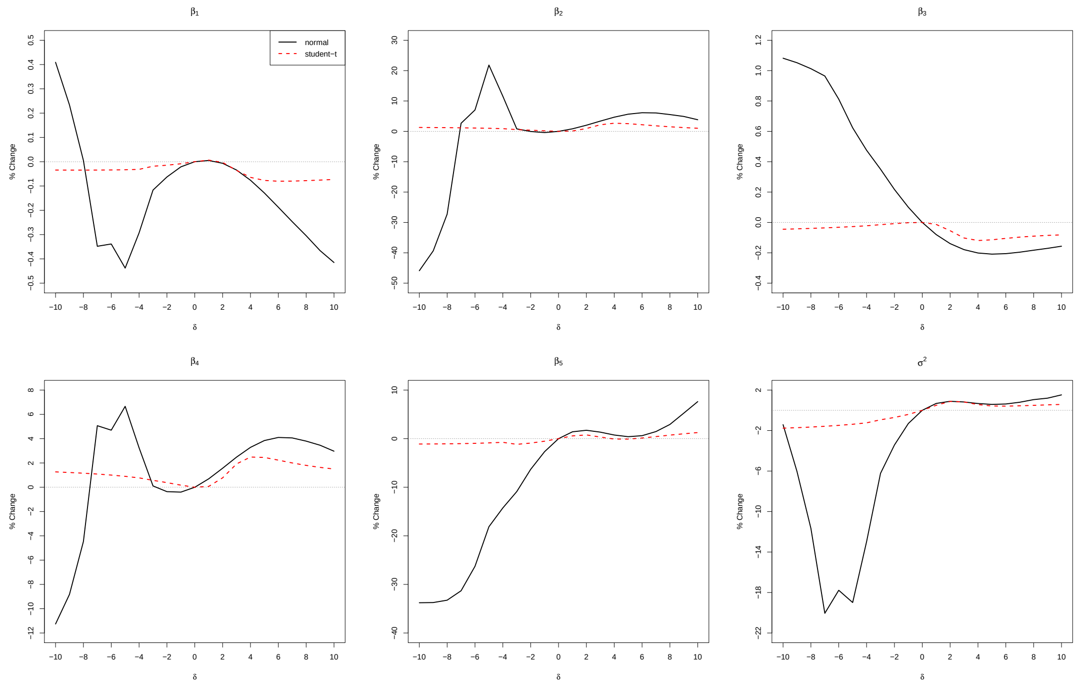


Fig. 6. AIEDRP data. Relative changes on the ML estimates of $\boldsymbol{\theta}$ from the N-NLMEC (solid line) and the t-NLMEC (dashed line) for different contaminations δ.

the effect of the level of censoring in the estimation. This way, we have 4 different simulation settings with 100 simulated datasets for each setting. Once the simulated data is generated, we fit the LMEC model assuming normal and Student-t distributions. For each of the simulations, we fit the model given in (30) assuming normal and Student-t distributions. For each simulation, the parameters estimation as well as AIC and BIC were recorded. Table 3 presents the summary statistics for β (the fixed-effects parameters) assuming normal and Student-t distributions for the 4 censoring patterns. In the Table, MC Mean denotes the arithmetic average of the 100 estimates given by $\sum_{j=1}^{100}\hat{\gamma}_j/100$ and MC Sd is the arithmetic average of the 100 posterior standard deviations given by $\sum_{j=1}^{100} sd(\hat{\gamma}_j)/100$, where $\gamma = \beta_1, \beta_2$ or σ^2. In addition, we also estimate the MC coverage of 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 3, we observe that the Student-t distribution over perform the normal distribution at all levels of censoring. Figure 7 shows that for the normal distribution there is a strong increase of the bias (the deviations of the parameter estimates from the true value) as well as the mean square error (MSE). Clearly, the Student-t model shows much less

Table 3. Monte Carlo results based on 100 simulated Student-t samples. MC mean, MC Sd (in parênteses) and MC Coverage are the respective mean estimates, standard deviations and coverage proportion average from fitting LMEC with Student-t and normal assumptions with different settings of censoring proportions. 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.

			Simulated Student-t data				
Censoring	Fit		β_1	β_2	σ^2	MC AIC	MC BIC
5%	Normal	MC Mean	-2.839	-0.179	0.285	604.261	626.484
		IM Sd	0.068	0.010			
		MC Sd	0.065	(0.006)	(0.072)		
		MC Coverage	98%	99%			
	Student-t	MC Mean	-2.831	-0.180	0.154	554.302	576.525
		IM Sd	0.055	0.008			
		MC Sd	(0.052)	(0.005)	(0.023)		
		MC Coverage	95%	100%			
10%	Normal	MC Mean	-2.822	-0.180	0.281	569.744	591.966
		IM Sd	0.070	0.010			
		MC Sd	(0.061)	(0.006)	(0.078)		
		MC Coverage	99%	99%			
	Student-t	MC Mean	-2.830	-0.179	0.150	526.334	548.557
		IM Sd	0.057	0.008			
		MC Sd	(0.059)	(0.006)	(0.024)		
		MC Coverage	97%	100%			
20%	Normal	MC Mean	-2.824	-0.180	0.270	505.704	527.927
		IM Sd	0.079	0.013			
		MC Sd	(0.076)	(0.009)	(0.073)		
		MC Coverage	97%	99%			
	Student-t	MC Mean	-2.832	-0.180	0.151	474.053	496.276
		IM Sd	0.068	0.011			
		MC Sd	(0.063)	(0.007)	(0.031)		
		MC Coverage	100%	99%			
50%	Normal	MC Mean	-2.810	-0.183	0.285	407.693	429.916
		IM Sd	0.090	0.016			
		MC Sd	(0.088)	(0.012)	(0.072)		
		MC Coverage	98%	99%			
	Student-t	MC Mean	-2.840	-0.178	0.154	387.582	409.805
		IM Sd	0.081	0.015			
		MC Sd	(0.066)	(0.007)	(0.023)		
		MC Coverage	98%	100%			

bias and thus more precise estimations. Therefore, models with heavier tails than normal produce more accurate estimates in the context of censored data; the degree and direction of the bias in fixed effects depends both on the relative proportions of censoring as well as model assumption. Observe that from Table 3 $\hat{\sigma}^2$ for the normal distribution is almost twice the true σ^2. This is due to the fact that in the normal scenario σ^2 represents the variance and therefore should be compared with $\frac{\nu}{\nu-2}\sigma^2$, which is 0.30. Notice also that, the Student-t model has a smaller confidence interval due to the smaller standard deviation but its coverage is slightly better than the normal method. This fact provides (once again) that the estimation of the Student-t method is more robust when dealing with censored data. Table 3 also provides the average values of the approximate standard deviations of the EM estimates obtained through the information-based method described in Subsection 3.4 (IM Sd) and the Monte Carlo standard deviation (Mc Sd) for the parameters. As we can see, the estimation method of the standard deviation provides relatively close results for the normal and Student-t methods, showing that the proposed asymptotic approximation for the variances of the fixed effects is reliable.

We also present the arithmetic average (MC AIC and MC BIC) of the model comparison criterions mentioned earlier. All the measures strongly favored the Student-t model, demonstrating the ability of these measures to detect an obvious departure from normality. The % of samples when these criteria chooses the t-LMEC also remains high.

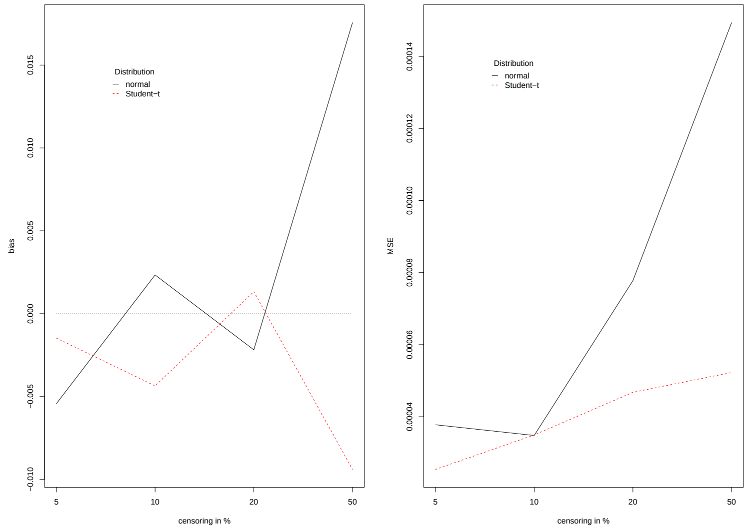
7.2. The nonlinear case

As in the linear case we fix the censoring proportion as presented in Section 7.1 and also generated 100 simulated data sets. Following Vaida and Liu (2009), to study the nonlinear regression, we consider the following nonlinear mixed model:

$$y_{ij} = \alpha_{1i} + \frac{\alpha_2}{(1+\exp((t_{ij}-\alpha_3)/\alpha_{4i}))} + \epsilon_{ij}, \quad i = 1,\ldots,100, \quad j = 1,\ldots,10, \tag{31}$$

where $(b_{1i}, b_{2i}) \overset{\text{iid.}}{\sim} t_2(\mathbf{0}, \mathbf{D}, \nu)$ and $\epsilon_{ij} \sim t(0, \sigma^2, \nu)$. We reparametrize $\beta_{1i} = \log(\alpha_{1i}) = \beta_1 + b_{1i}$; $\beta_k = \log(\alpha_k), k = 2, 3$, $\alpha_{4i} = \beta_4 + b_{2i}$ and in addition, we set $t_{ij} = (1, 10, 20, 30, 40, 50, 60, 70, 80, 90)$, $\boldsymbol{\beta}^\top = (\beta_1, \beta_2, \beta_3, \beta_4) = (1.6094, 0.6931, 3.8067, 2.3026)$, $\mathbf{D} = \begin{bmatrix} 0.0025 & -0.0010 \\ -0.0010 & 0.0100 \end{bmatrix}$, $\sigma^2 = 0.55$ and $\nu = 4$.

We fit the NLMEC model (31) assuming normal and Student-t distributions. For each of the simulations, we fit the reparameterized model given in (31) assuming normal and Student-t distributions. The model selection criterion AIC and BIC as well as the pa-



(a) Bias of β_1 (b) MSE of β_1

Fig. 7. (a) Represents the bias of β_1 in comparison with the true value for the normal and Student-t models for the 4 censoring patterns (5%, 10%, 20%, 50%)in the LMEC setup. (b) Presents the Mean Square Error (MSE) for β_1 for the normal and Student-t models.

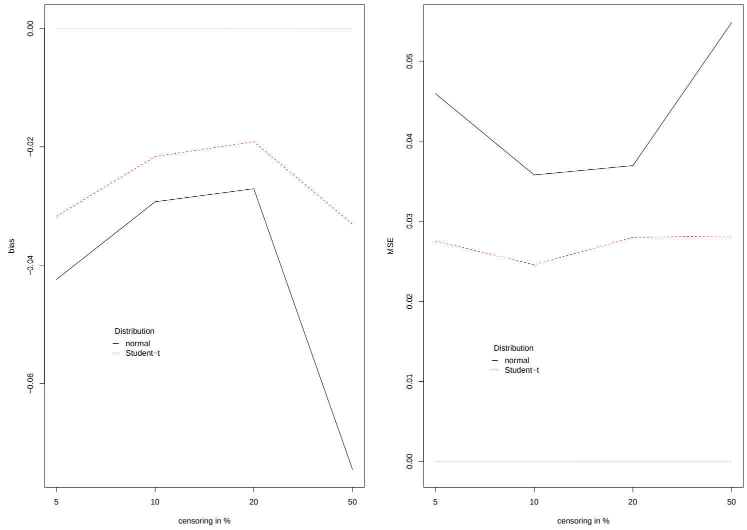
rameters estimation were recorded for each simulation. For the 4 censoring patterns, the summary statistics for β (the fixed-effects parameters) are presented in Table 4 assuming normal and Student-t distributions.

From Table 4, we observe that for all levels of censoring the Student-t distribution performs better than the normal distribution and have a small standard deviation in the estimates providing more accurate estimation. The arithmetic average (MC AIC and MC BIC) of the model comparison criteria are also presented and strongly favors the Student-t model in comparison to the normal model. This, reinforce that these measures are capable of detecting departures from normality. Like in the linear case, we have that the estimates $\hat{\sigma}^2$ of σ^2 for the normal distribution must be compared with $\frac{\nu}{\nu-2}\sigma^2$, which now is 1.10. As in the linear setup we can see that the Student-t model continues to have smaller confidence interval with a usually bigger coverage of the parameters. This is a strong evidence of the robustness in estimation of the Student-t method. Again, as observed in the linear case the IM Sd and MC Sd for the nonlinear regression provides close results for both models (normal and Student-t). This emphasize that the estimation of the standard error provided by the proposed asymptotic approximation of the fixed effects (Equation 22) is reliable.

In Figure 8 we represent the bias and MSE for the parameter estimates of β_4 for the

Table 4. Monte Carlo results based on 100 simulated Student-t samples. MC mean, MC Sd (in parenthesis) and MC Coverage are the respective mean estimates, standard deviations and coverage proportion average from fitting NLMEC with Student-t and normal assumptions with different settings of censoring proportions. IM Sd are the average values of the approximate standard error obtained through the information-based method. MC AIC and MC BIC are the arithmetic average of the respective model comparison measures.

			Simulated Student-t data						
Censoring	Fit		β_1	β_2	β_3	β_4	σ^2	MC AIC	MC BIC
5%	Normal	MC Mean	1.627	0.642	3.796	2.205	0.967	2865.279	2904.541
		IM Sd	0.017	0.068	0.041	0.191			
		MC Sd	(0.016)	(0.073)	(0.043)	(0.192)	(0.146)		
		MC coverage	81%	87%	96%	95%			
	Student-t	MC Mean	1.615	0.667	3.805	2.230	0.642	2654.928	2694.190
		IM Sd	0.015	0.058	0.035	0.161			
		MC Sd	(0.012)	(0.056)	(0.031)	(0.150)	(0.060)		
		MC coverage	96%	93%	99%	95%			
10%	Normal	MC Mean	1.623	0.657	3.801	2.235	0.970	2815.475	2854.737
		IM Sd	0.018	0.070	0.042	0.191			
		MC Sd	(0.017)	(0.069)	(0.046)	(0.178)	(0.141)		
		MC coverage	86%	88%	92%	95%			
	Student-t	MC Mean	1.613	0.676	3.803	2.253	0.629	2608.471	2647.733
		IM Sd	0.015	0.059	0.035	0.160			
		MC Sd	(0.014)	(0.057)	(0.036)	(0.150)	(0.057)		
		MC coverage	94%	94%	95%	97%			
20%	Normal	MC Mean	1.616	0.683	3.806	2.240	0.975	2705.762	2494.963
		IM Sd	0.019	0.070	0.042	0.190			
		MC Sd	(0.016)	(0.069)	(0.042)	(0.183)	(0.145)		
		MC coverage	95%	95%	98%	96%			
	Student-t	MC Mean	1.616	0.678	3.797	2.259	0.579	2494.963	2534.225
		IM Sd	0.015	0.059	0.035	0.157			
		MC Sd	(0.015)	(0.060)	(0.032)	(0.162)	(0.044)		
		MC coverage	89%	92%	99%	95%			
50%	Normal	MC Mean	1.614	0.684	3.781	2.131	0.978	1982.382	2021.644
		IM Sd	0.022	0.073	0.043	0.208			
		MC Sd	(0.023)	(0.069)	(0.045)	(0.160)	(0.186)		
		MC coverage	94%	95%	90%	93%			
	Student-t	MC Mean	1.624	0.650	3.789	2.226	0.546	1879.266	1918.528
		IM Sd	0.022	0.075	0.041	0.187			
		MC Sd	(0.016)	(0.066)	(0.040)	(0.151)	(0.038)		
		MC coverage	90%	93%	95%	95%			



(a) Bias of β_4 (b) MSE of β_4

Fig. 8. (a) Represents the bias of β_4 in comparison with the true value for the normal and Student-t models for the 4 censoring patterns (5%, 10%, 20%, 50%) in the NLMEC setup. (b) Presents the Mean Square Error (MSE) for β_4 for the normal and Student-t models.

normal and Student-t distributions. It is clear that the normal model has a much bigger bias and MSE than the Student-t model. Therefore, for censored data the Student-t model is more robust, providing more accurate estimations when the data has departures from the normality assumption. Although Figure 8 only presents the results for the estimates of β_4 a similar pattern was observed for all the other parameters.

8. Conclusions

We have proposed a robust approach to linear and nonlinear mixed effects models with censored observation based on the multivariate-t distribution, called the t-LMEC/t-NLMEC. It offers a great deal of flexibility in dealing with longitudinal data in the presence of outliers. A novel ECM algorithm to obtain approximated MLEs is developed by exploring the statistical properties of the multivariate truncated Student-t distribution. Our proposed algorithm has a closed-form expression for the E-step, based on formulas for the mean and variance of the truncated Student-t distribution. Thus, the proposed methodologies allow the practitioner to fit longitudinal data in a broad variety of considerations. For NLMEC, the analysis is computationally feasible through approximating the t-NLMEC for a multivariate t distribution with specified parameters. We apply our methodology to two recent AIDS studies (freely downloadable from R) as well as simulated data to illustrate how the procedures can be used to evaluate model assumptions, identify outliers, and obtain robust parameter estimates. From these results it is encouraging that the use of t-LMEC/t-NLMEC models offer better fitting, protection against outliers and more precise inferences than the usual normal counterpart.

It may be worthwhile comparing our results in NLMEC with other methods such as the classical Monte Carlo EM algorithm or the stochastic version of the EM algorithm (SAEM) which is beyond the scope of this paper. These issues will be considered in a separate future work. We conjecture that the methodology presented in this paper should yield satisfactory results in other areas where multivariate data appears frequently, for instance, measurement error models, correlated binary regression, dynamics linear models, spatially censored data, etc., at expense of moderate complexity of implementation. Finally, the proposed EM algorithm has been coded and implemented in the R package (R Development Core Team, 2009) and is available from the authors upon request.

Acknowledgements

We thank Dr. Tsung Lin for supplying his R-code on truncated Student-t distribution which has been proved to be very helpful in implementing our procedure. V. H. Lachos acknowledges support of the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq-Brazil) and the Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP-Brazil).

Appendix

Appendix A: The EM algorithm

We include here the derivation of the EM equations (18) - (20). Recall that the vector of parameters to be estimated is $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \sigma^2, \boldsymbol{\alpha})$ and that $\mathbf{y} = (\mathbf{y}_1^\top, \ldots, \mathbf{y}_n^\top)^\top$, $\mathbf{b} = (\mathbf{b}_1^\top, \ldots, \mathbf{b}_n^\top)^\top$, $\mathbf{u} = (u_1, \ldots, u_n)^\top$, $\mathbf{Q} = vec(\mathbf{Q}_1, \ldots, \mathbf{Q}_n)$ and $\mathbf{C} = vec(\mathbf{C}_1, \ldots, \mathbf{C}_n)$, such that we observe $(\mathbf{Q}_i, \mathbf{C}_i)$ for the ith subject. In their estimation procedure, $\mathbf{b}$, $\mathbf{Q}$ and $\mathbf{C}$ are treated as hypothetical missing data, and augmented with the observed data set $\mathbf{y}_c = (\mathbf{C}^\top, \mathbf{Q}^\top, \mathbf{y}^\top, \mathbf{b}^\top, \mathbf{u}^\top)^\top$.

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

The complete log-likelihood is given by

$$\begin{aligned}\ell_c(\boldsymbol{\theta}|\mathbf{y}_c) &= \log(L(\mathbf{y}_c|\boldsymbol{\theta})) = C + \sum_{i=1}^{n}\Big\{h(u_i|\nu) - \frac{1}{2}\left[n_i \log\sigma^2 + \log|\mathbf{D}| + u_i\mathbf{b}_i^\top\mathbf{D}^{-1}\mathbf{b}_i\right.\\ &\left.+\frac{u_i}{\sigma^2}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{Z}_i\mathbf{b}_i)^\top(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{Z}_i\mathbf{b}_i)\right]\Big\},\end{aligned}$$

where C is a constant that is independent of the parameter vector $\boldsymbol{\theta}$ and $h(u_i|\nu)$ is a density of a $Gamma(\nu/2,\nu/2)$. The EM function is given by

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

So we have that,

$$\begin{aligned}Q(\boldsymbol{\theta}|\boldsymbol{\theta}^*) &= C^* - \frac{1}{2}\sum_{i=1}^{n}\left[n_i\log\sigma^2 + \log|\mathbf{D}| + \operatorname{tr}\left(E[u_i\mathbf{b}_i\mathbf{b}_i^\top|\mathbf{Q}_i,\mathbf{C}_i,\boldsymbol{\theta}^*]\mathbf{D}^{-1}\right)\right.\\ &\left.+E\left[\frac{u_i}{\sigma^2}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{Z}_i\mathbf{b}_i)^\top(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{Z}_i\mathbf{b}_i)|\mathbf{Q}_i,\mathbf{C}_i,\boldsymbol{\theta}^*\right]\right],\end{aligned}$$

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

In order to introduce some important results, we establish the following lemma. The 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,*

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

where $\boldsymbol{\Lambda} = (\boldsymbol{\Omega}^{-1} + \mathbf{A}^\top\boldsymbol{\Sigma}^{-1}\mathbf{A})^{-1}$.

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

$$\mathbf{y}_i \overset{\text{ind.}}{\sim} Tt_{n_i}(\mathbf{X}_i\boldsymbol{\beta},\boldsymbol{\Sigma}_i,\nu),$$

$$E(u_i|\mathbf{y}_i) = \frac{\nu+n_i}{\nu+\delta},$$

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

$$\mathbf{b}_i|\mathbf{y}_i,u_i \overset{\text{ind.}}{\sim} N_q\left(\frac{u_i}{\sigma^2}\left(u_i\mathbf{D}^{-1}+\frac{u_i}{\sigma^2}\mathbf{Z}_i^\top\mathbf{Z}_i\right)^{-1}\mathbf{Z}_i^\top(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta}),\left(u_i\mathbf{D}^{-1}+\frac{u_i}{\sigma^2}\mathbf{Z}_i^\top\mathbf{Z}_i\right)^{-1}\right),$$

$$\mathbf{b}_i|\mathbf{y}_i,u_i \overset{\text{ind.}}{\sim} N_q\left(\boldsymbol{\varphi}_i(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta}),\frac{\sigma^2}{u_i}\boldsymbol{\Lambda}_i\right),$$

with $\boldsymbol{\Lambda}_i=(\sigma^2\mathbf{D}^{-1}+\mathbf{Z}_i^\top\mathbf{Z}_i)^{-1}$ and $\boldsymbol{\varphi}_i=\boldsymbol{\Lambda}_i\mathbf{Z}_i^\top$. Using the propositions (1)-(3) we compute the following expectation terms:

$$\begin{aligned}
\widehat{u\mathbf{y}_i} &= E\{u_i\mathbf{y}_i|\mathbf{Q}_i,\mathbf{C}_i,\boldsymbol{\theta}^*\}=E_{\mathbf{y}_i|\mathbf{Q}_i,\mathbf{C}_i}\{E_{u_i|\mathbf{y}_i}[E_{\mathbf{b}_i|\mathbf{y}_i,u_i}(u\mathbf{y}_i)]\}\\
&= E_{\mathbf{y}_i|\mathbf{Q}_i,\mathbf{C}_i}\left[E_{u_i|\mathbf{y}_i}(u_i\mathbf{y}_i)\right]=E_{\mathbf{y}_i|\mathbf{Q}_i,\mathbf{C}_i}\left(\frac{(\nu+n_i)}{(\nu+\delta)}\mathbf{y}_i\right)\\
&= \frac{T_{n_i}(a|\boldsymbol{\mu},\boldsymbol{\Sigma}^*,\nu+2)}{T_{n_i}(a|\boldsymbol{\mu},\boldsymbol{\Sigma},\nu)}E\{\mathbf{W}_i\}.
\end{aligned}$$

$$\begin{aligned}
\widehat{u\mathbf{y}_i^2} &= E\{u_i\mathbf{y}_i\mathbf{y}_i^\top|\mathbf{Q}_i,\mathbf{C}_i,\boldsymbol{\theta}^*\}=E_{\mathbf{y}_i|\mathbf{Q}_i,\mathbf{C}_i}\{E_{u_i|\mathbf{y}_i}[E_{\mathbf{b}_i|\mathbf{y}_i,u_i}(u\mathbf{y}_i\mathbf{y}_i^\top)]\}\\
&= E_{\mathbf{y}_i|\mathbf{Q}_i,\mathbf{C}_i}\left[E_{u_i|\mathbf{y}_i}(u_i\mathbf{y}_i\mathbf{y}_i^\top)\right]=E_{\mathbf{y}_i|\mathbf{Q}_i,\mathbf{C}_i}\left(\frac{(\nu+n_i)}{(\nu+\delta)}\mathbf{y}_i\mathbf{y}_i^\top\right)\\
&= \frac{T_{n_i}(a|\boldsymbol{\mu},\boldsymbol{\Sigma}^*,\nu+2)}{T_{n_i}(a|\boldsymbol{\mu},\boldsymbol{\Sigma},\nu)}E\{\mathbf{W}_i\mathbf{W}_i^\top\}.
\end{aligned}$$

$$\begin{aligned}
\widehat{u_i} &= E\{u_i|\mathbf{Q}_i,\mathbf{C}_i,\boldsymbol{\theta}^*\}=E_{\mathbf{y}_i|\mathbf{Q}_i,\mathbf{C}_i}\{E_{u_i|\mathbf{y}_i}[E_{\mathbf{b}_i|\mathbf{y}_i,u_i}(u_i)]\}\\
&= E_{\mathbf{y}_i|\mathbf{Q}_i,\mathbf{C}_i}\left[E_{u_i|\mathbf{y}_i}(u_i)\right]=E_{\mathbf{y}_i|\mathbf{Q}_i,\mathbf{C}_i}\left(\frac{(\nu+n_i)}{(\nu+\delta)}\right)\\
&= \frac{T_{n_i}(a|\boldsymbol{\mu},\boldsymbol{\Sigma}^*,\nu+2)}{T_{n_i}(a|\boldsymbol{\mu},\boldsymbol{\Sigma},\nu)}E\{\mathbf{W}_i^0\}=\frac{T_{n_i}(a|\boldsymbol{\mu},\boldsymbol{\Sigma}^*,\nu+2)}{T_{n_i}(a|\boldsymbol{\mu},\boldsymbol{\Sigma},\nu)}.
\end{aligned}$$

$$\begin{aligned}
\widehat{u\mathbf{b}_i} &= E\{u_i\mathbf{b}_i|\mathbf{Q}_i,\mathbf{C}_i,\boldsymbol{\theta}^*\}=E_{\mathbf{y}_i|\mathbf{Q}_i,\mathbf{C}_i}\{E_{u_i|\mathbf{y}_i}[E_{\mathbf{b}_i|\mathbf{y}_i,u_i}(u_i\mathbf{b}_i)]\}\\
&= E_{\mathbf{y}_i|\mathbf{Q}_i,\mathbf{C}_i}\{E_{u_i|\mathbf{y}_i}[u_iE_{\mathbf{b}_i|\mathbf{y}_i,u_i}(\mathbf{b}_i)]\}\\
&= E_{\mathbf{y}_i|\mathbf{Q}_i,\mathbf{C}_i}\{E_{u_i|\mathbf{y}_i}\left[u_i\boldsymbol{\varphi}_i(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta})\right]\}\\
&= E_{\mathbf{y}_i|\mathbf{Q}_i,\mathbf{C}_i}\left[\boldsymbol{\varphi}_i(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta})E_{u_i|\mathbf{y}_i}(u_i)\right]\\
&= E_{\mathbf{y}_i|\mathbf{Q}_i,\mathbf{C}_i}\left\{\boldsymbol{\varphi}_i\left[\left(\frac{(\nu+n_i)}{(\nu+\delta)}\mathbf{y}_i\right)-\mathbf{X}_i\boldsymbol{\beta}\left(\frac{(\nu+n_i)}{(\nu+\delta)}\right)\right]\right\}\\
&= \boldsymbol{\varphi}_i\left[E_{\mathbf{y}_i|\mathbf{Q}_i,\mathbf{C}_i}\left(\frac{(\nu+n_i)}{(\nu+\delta)}\mathbf{y}_i\right)-\mathbf{X}_i\boldsymbol{\beta}E_{\mathbf{y}_i|\mathbf{Q}_i,\mathbf{C}_i}\left(\frac{(\nu+n_i)}{(\nu+\delta)}\right)\right]\\
&= \boldsymbol{\varphi}_i\left[\frac{T_{n_i}(a|\boldsymbol{\mu},\boldsymbol{\Sigma}^*,\nu+2)}{T_{n_i}(a|\boldsymbol{\mu},\boldsymbol{\Sigma},\nu)}E\{\mathbf{W}_i\}-\mathbf{X}_i\boldsymbol{\beta}\frac{T_{n_i}(a|\boldsymbol{\mu},\boldsymbol{\Sigma}^*,\nu+2)}{T_{n_i}(a|\boldsymbol{\mu},\boldsymbol{\Sigma},\nu)}\right]\\
&= \boldsymbol{\varphi}_i\left[\widehat{u\mathbf{y}_i}-\mathbf{X}_i\boldsymbol{\beta}\widehat{u_i}\right].
\end{aligned}$$

$$
\begin{aligned}
\widehat{u\mathbf{b}^2_i} &= E\{u_i\mathbf{b}_i\mathbf{b}_i^\top|\mathbf{Q}_i,\mathbf{C}_i,\boldsymbol{\theta}^*\} = E_{\mathbf{y}_i|\mathbf{Q}_i,\mathbf{C}_i}\{E_{u_i|\mathbf{y}_i}[E_{\mathbf{b}_i|\mathbf{y}_i,u_i}(u_i\mathbf{b}_i\mathbf{b}_i^\top)]\} \\
&= E_{\mathbf{y}_i|\mathbf{Q}_i,\mathbf{C}_i}\{E_{u_i|\mathbf{y}_i}[u_iE_{\mathbf{b}_i|\mathbf{y}_i,u_i}(\mathbf{b}_i\mathbf{b}_i^\top)]\} \\
&= E_{\mathbf{y}_i|\mathbf{Q}_i,\mathbf{C}_i}\left\{E_{u_i|\mathbf{y}_i}\left[u_i\left(\boldsymbol{\Lambda}_i(u_i^{-1}\sigma^2+\mathbf{Z}_i^\top(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta})(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta})^\top)\boldsymbol{\varphi}_i^\top\right)\right]\right\} \\
&= \boldsymbol{\Lambda}_i\sigma^2+E_{\mathbf{y}_i|\mathbf{Q}_i,\mathbf{C}_i}\left[E_{u_i|\mathbf{y}_i}(u_i)\left(\boldsymbol{\varphi}_i(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta})(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta})^\top\boldsymbol{\varphi}_i^\top\right)\right] \\
&= \boldsymbol{\Lambda}_i\sigma^2+\boldsymbol{\varphi}_i\left\{E_{\mathbf{y}_i|\mathbf{Q}_i,\mathbf{C}_i}\left(\frac{(\nu+n_i)}{(\nu+\delta)}\mathbf{y}_i\mathbf{y}_i^\top\right)-E_{\mathbf{y}_i|\mathbf{Q}_i,\mathbf{C}_i}\left(\frac{(\nu+n_i)}{(\nu+\delta)}\mathbf{y}_i\right)\boldsymbol{\beta}^\top\mathbf{X}_i^\top\right. \\
&\quad\left.-\mathbf{X}_i\boldsymbol{\beta}\left[E_{\mathbf{y}_i|\mathbf{Q}_i,\mathbf{C}_i}\left(\frac{(\nu+n_i)}{(\nu+\delta)}\mathbf{y}_i\right)\right]^\top+E_{\mathbf{y}_i|\mathbf{Q}_i,\mathbf{C}_i}\left(\frac{(\nu+n_i)}{(\nu+\delta)}\right)\mathbf{X}_i\boldsymbol{\beta}\boldsymbol{\beta}^\top\mathbf{X}_i^\top\right\}\boldsymbol{\varphi}_i^\top \\
&= \boldsymbol{\Lambda}_i\sigma^2+\boldsymbol{\varphi}_i\left\{\frac{T_{n_i}(a|\boldsymbol{\mu},\boldsymbol{\Sigma}^*,\nu+2)}{T_{n_i}(a|\boldsymbol{\mu},\boldsymbol{\Sigma},\nu)}E\{\mathbf{W}_i\mathbf{W}_i^\top\}-\frac{T_{n_i}(a|\boldsymbol{\mu},\boldsymbol{\Sigma}^*,\nu+2)}{T_{n_i}(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_{n_i}(a|\boldsymbol{\mu},\boldsymbol{\Sigma}^*,\nu+2)}{T_{n_i}(a|\boldsymbol{\mu},\boldsymbol{\Sigma},\nu)}E\{\mathbf{W}_i\}\right]^\top+\frac{T_{n_i}(a|\boldsymbol{\mu},\boldsymbol{\Sigma}^*,\nu+2)}{T_{n_i}(a|\boldsymbol{\mu},\boldsymbol{\Sigma},\nu)}\mathbf{X}_i\boldsymbol{\beta}\boldsymbol{\beta}^\top\mathbf{X}_i^\top\right\}\boldsymbol{\varphi}_i^\top \\
&= \boldsymbol{\Lambda}_i\sigma^2+\boldsymbol{\varphi}_i\left(\widehat{u\mathbf{y}^2_i}-\widehat{u\mathbf{y}_i}\boldsymbol{\beta}^\top\mathbf{X}_i^\top-\mathbf{X}_i\boldsymbol{\beta}\widehat{u\mathbf{y}_i}^\top+\widehat{u}_i\mathbf{X}_i\boldsymbol{\beta}\boldsymbol{\beta}^\top\mathbf{X}_i^\top\right)\boldsymbol{\varphi}_i^\top.
\end{aligned}
$$

$$
\begin{aligned}
\widehat{u\mathbf{y}\mathbf{b}_i} &= E\{u_i\mathbf{y}_i\mathbf{b}_i^\top|\mathbf{Q}_i,\mathbf{C}_i,\boldsymbol{\theta}^*\} = E_{\mathbf{y}_i|\mathbf{Q}_i,\mathbf{C}_i}\{E_{u_i|\mathbf{y}_i}[E_{\mathbf{b}_i|\mathbf{y}_i,u_i}(u_i\mathbf{y}_i\mathbf{b}_i^\top)]\} \\
&= E_{\mathbf{y}_i|\mathbf{Q}_i,\mathbf{C}_i}\{\mathbf{y}_iE_{u_i|\mathbf{y}_i}[u_iE_{\mathbf{b}_i|\mathbf{y}_i,u_i}(\mathbf{b}_i^\top)]\} \\
&= E_{\mathbf{y}_i|\mathbf{Q}_i,\mathbf{C}_i}\left\{\mathbf{y}_iE_{u_i|\mathbf{y}_i}\left[u_i(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta})^\top\boldsymbol{\varphi}_i^\top\right]\right\} \\
&= E_{\mathbf{y}_i|\mathbf{Q}_i,\mathbf{C}_i}\left[\mathbf{y}_iE_{u_i|\mathbf{y}_i}(u_i)(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta})^\top\boldsymbol{\varphi}_i^\top\right] \\
&= E_{\mathbf{y}_i|\mathbf{Q}_i,\mathbf{C}_i}\left\{\left[\left(\frac{(\nu+n_i)}{(\nu+\delta)}\mathbf{y}_i\mathbf{y}_i^\top\right)-\left(\frac{(\nu+n_i)}{(\nu+\delta)}\mathbf{y}_i\right)\boldsymbol{\beta}^\top\mathbf{X}_i^\top\right]\boldsymbol{\varphi}_i^\top\right\} \\
&= \left[E_{\mathbf{y}_i|\mathbf{Q}_i,\mathbf{C}_i}\left(\frac{(\nu+n_i)}{(\nu+\delta)}\mathbf{y}_i\mathbf{y}_i^\top\right)-E_{\mathbf{y}_i|\mathbf{Q}_i,\mathbf{C}_i}\left(\frac{(\nu+n_i)}{(\nu+\delta)}\mathbf{y}_i\right)\boldsymbol{\beta}^\top\mathbf{X}_i^\top\right]\boldsymbol{\varphi}_i^\top \\
&= \left[\frac{T_{n_i}(a|\boldsymbol{\mu},\boldsymbol{\Sigma}^*,\nu+2)}{T_{n_i}(a|\boldsymbol{\mu},\boldsymbol{\Sigma},\nu)}E\{\mathbf{W}_i\}-\frac{T_{n_i}(a|\boldsymbol{\mu},\boldsymbol{\Sigma}^*,\nu+2)}{T_{n_i}(a|\boldsymbol{\mu},\boldsymbol{\Sigma},\nu)}E\{\mathbf{W}_i\mathbf{W}_i^\top\}\boldsymbol{\beta}^\top\mathbf{X}_i^\top\right]\boldsymbol{\varphi}_i^\top \\
&= \left[\widehat{u\mathbf{y}^2_i}-\widehat{u\mathbf{y}_i}\boldsymbol{\beta}^\top\mathbf{X}_i^\top\right]\boldsymbol{\varphi}_i^\top.
\end{aligned}
$$

Replacing the expectation in $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^*)$

$$
Q(\boldsymbol{\theta}|\boldsymbol{\theta}^*) = C^*-\frac{1}{2}\sum_{i=1}^{n}\left[n_i\log\sigma^2+\log|\mathbf{D}|+\mathrm{tr}\left(\widehat{u\mathbf{b}^2_i}\mathbf{D}^{-1}\right)+\frac{A_i}{\sigma^2}\right],
$$

where

$$
\begin{aligned}
A_i &= \mathrm{tr}(\widehat{u\mathbf{y}^2_i})-\widehat{u\mathbf{y}_i}^\top\mathbf{X}_i\boldsymbol{\beta}-\mathrm{tr}(\widehat{u\mathbf{y}\mathbf{b}_i}^\top\mathbf{Z}_i)-\boldsymbol{\beta}^\top\mathbf{X}_i^\top\widehat{u\mathbf{y}_i}+\boldsymbol{\beta}^\top\mathbf{X}_i^\top\widehat{u}_i\mathbf{X}_i\boldsymbol{\beta}, \\
&\quad+\boldsymbol{\beta}^\top\mathbf{X}_i^\top\mathbf{Z}_i\widehat{u\mathbf{b}_i}-\mathrm{tr}(\widehat{u\mathbf{y}\mathbf{b}_i}\mathbf{Z}_i^\top)+\widehat{u\mathbf{b}_i}^\top\mathbf{Z}_i^\top\mathbf{X}_i\boldsymbol{\beta}+\mathrm{tr}(\widehat{u\mathbf{b}^2_i}\mathbf{Z}_i^\top\mathbf{Z}_i).
\end{aligned}
$$

The differential with respect to $\boldsymbol{\beta}$, σ^2 and $\mathbf{D}$ are

$$\begin{aligned}
\frac{\partial Q(\boldsymbol{\theta}|\boldsymbol{\theta}^*)}{\partial \boldsymbol{\beta}} &= -\frac{1}{\sigma^2}\sum_{i=1}^{n} -\mathbf{X}_i^\top (\widehat{u\mathbf{y}_i} - \widehat{u}_i \mathbf{X}_i \boldsymbol{\beta} - \mathbf{Z}_i \widehat{u\mathbf{b}_i}),\\
\frac{\partial Q(\boldsymbol{\theta}|\boldsymbol{\theta}^*)}{\partial \sigma^2} &= -\frac{1}{2}\sum_{i=1}^{n}\left[\frac{n_i}{\sigma^2} - \frac{A_i}{(\sigma^2)^2}\right],\\
\frac{\partial Q(\boldsymbol{\theta}|\boldsymbol{\theta}^*)}{\partial \mathbf{D}^{-1}} &= -\frac{n}{\sigma^2}(-2\mathbf{D} + diag(\mathbf{D})) - \frac{1}{2}\sum_{i=1}^{n}\left(\widehat{u\mathbf{y}_i} + \widehat{u\mathbf{y}_i}^\top - diag(\widehat{u\mathbf{b}_i^2})\right).
\end{aligned}$$

The solution of $\dfrac{\partial Q(\boldsymbol{\theta}|\boldsymbol{\theta}^*)}{\partial \boldsymbol{\beta}} = 0$ is

$$\widehat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{n} \mathbf{X}_i^\top \widehat{u}_i \mathbf{X}_i\right)^{-1}\left[\sum_{i=1}^{n} \mathbf{X}_i (\widehat{u\mathbf{y}_i} - \mathbf{Z}_i \widehat{u\mathbf{b}_i})\right].$$

The solution of $\dfrac{\partial Q(\boldsymbol{\theta}|\boldsymbol{\theta}^*)}{\partial \sigma^2} = 0$ is

$$\widehat{\sigma^2} = \frac{\sum_{i=1}^{n} A_i}{\sum_{i=1}^{n} n_i}.$$

For unstructured $\mathbf{D}$, the solution of $\dfrac{\partial Q(\boldsymbol{\theta}|\boldsymbol{\theta}^*)}{\partial \mathbf{D}^{-1}} = 0$ for all $\mathbf{D}$ is The solution of $\dfrac{\partial Q(\boldsymbol{\theta}|\boldsymbol{\theta}^*)}{\partial \sigma^2} = 0$ is

$$\widehat{\mathbf{D}} = \sum_{i=1}^{n} \frac{\widehat{u\mathbf{b}_i^2}}{n}.$$

Appendix B: The expected information matrix of the fixed effects

In this Appendix we derived the expected information matrix for the fixed effects. Thus, using the method given by McLachlan and Krishnan (1996), we have that

$$I(\boldsymbol{\beta}; \mathbf{y}) = I_c(\boldsymbol{\beta}; \mathbf{y}) + I_m(\boldsymbol{\beta}; \mathbf{y}),$$

where $I(\boldsymbol{\beta}; \mathbf{y})$ is the information matrix about $\boldsymbol{\beta}$ in the observed data $\mathbf{y}$, $I_c(\boldsymbol{\beta}; \mathbf{y})$ is the conditional expectation of the complete-data information matrix, and $I_m(\boldsymbol{\beta}; \mathbf{y})$ is the missing information matrix.

The missing data information $I_m(\boldsymbol{\beta}; \mathbf{y})$ can be expressed as

$$I_m(\boldsymbol{\beta}; \mathbf{y}) = \sum_{i=1}^{n} Var\left\{S_c(\mathbf{y}; \boldsymbol{\beta})|\mathbf{Q}_i, \mathbf{C}_i\right\},$$

where $S_c(\mathbf{y};\boldsymbol{\beta}) = \frac{\partial log L_c(\mathbf{y};\boldsymbol{\theta})}{\partial\boldsymbol{\beta}}$ is the gradient vector of the complete-data log likelihood function. So, we have that

$$\begin{aligned} I_m(\boldsymbol{\beta};\mathbf{y}) &= \sum_{i=1}^{n} Var\left(\left(\frac{\nu+n_i}{\nu+\delta}\right)\mathbf{X}_i^\top\boldsymbol{\Sigma}_i^{-1}(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta})|\mathbf{Q}_i,\mathbf{C}_i\right) \\ &= \mathbf{X}_i^\top\boldsymbol{\Sigma}_i^{-1}\left\{Var\left(\left(\frac{\nu+n_i}{\nu+\delta}\right)(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta})|\mathbf{Q}_i,\mathbf{C}_i\right)\right\}\boldsymbol{\Sigma}_i^{-1}\mathbf{X}_i. \end{aligned}$$

Now by using the results given in Lange et al. (1989) (Appendix B), the expected (complete-data) information matrix is given by

$$I_c(\boldsymbol{\beta};\mathbf{y}) = \sum_{i=1}^{n}\frac{\nu+n_i}{\nu+n_i+2}\mathbf{X}_i^\top\boldsymbol{\Sigma}_i^{-1}\mathbf{X}_i.$$

It follows that the observed information matrix is given by

$$I(\boldsymbol{\beta};\mathbf{y}) = \sum_{i=1}^{n}\frac{\nu+n_i}{\nu+n_i+2}\mathbf{X}_i^\top\boldsymbol{\Sigma}_i^{-1}\mathbf{X}_i - \sum_{i=1}^{n}\mathbf{X}_i^\top\boldsymbol{\Sigma}_i^{-1}\mathbf{B}_i\boldsymbol{\Sigma}_i^{-1}\mathbf{X}_i,$$

where $\mathbf{B}_i = Var\left\{\frac{\nu+n_i}{\nu+Q(\mathbf{y}_i)}(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta})|\mathbf{Q}_i,\mathbf{C}_i\right\}$, with $\mathbf{y}_i \sim Tt_{n_i}(\mathbf{X}_i\boldsymbol{\beta},\boldsymbol{\Sigma}_i,\nu;\mathbb{A}_i)$.

Appendix C: More general linear mixed effects models

Heteroscedastics Error

We include here the derivation of the equations (23) - (25).

For the general linear mixed effects model represented in (23) - (25), with $\boldsymbol{\Omega}_{n_i}$ with a first order autoregressive structure (AR(1)), the complete log-likelihood is given by

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

where C is a constant that is independent of the parameter vector $\boldsymbol{\theta}$ and $h(u_i|\nu)$ is a density of a $Gamma(\nu/2,\nu/2)$. The EM function is given by

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^*) &= C^* - \frac{1}{2}\sum_{i=1}^{n}\left\{n_i\log\sigma^2 + \log|\boldsymbol{\Omega}_{n_i}| + \log|\mathbf{D}| + \mathrm{tr}\left(E[u_i\mathbf{b}_i\mathbf{b}_i^\top|\mathbf{Q}_i,\mathbf{C}_i,\boldsymbol{\theta}^*]\mathbf{D}^{-1}\right)\right. \\ &\left. + E\left[\frac{u_i}{\sigma^2}(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta}-\mathbf{Z}_i\mathbf{b}_i)^\top\boldsymbol{\Omega}_{n_i}^{-1}(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta}-\mathbf{Z}_i\mathbf{b}_i)|\mathbf{Q}_i,\mathbf{C}_i,\boldsymbol{\theta}^*\right]\right\}, \end{aligned}$$

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

Let $\boldsymbol{\epsilon}_i = \mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{Z}_i\mathbf{b}_i$, so

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^*) &= C^* - \frac{1}{2}\sum_{i=1}^{n}\Big\{n_i \log\sigma^2 + \log|\boldsymbol{\Omega}_{n_i}| + \log|\mathbf{D}| + \text{tr}\left(E[u_i\mathbf{b}_i\mathbf{b}_i^\top|\mathbf{Q}_i, \mathbf{C}_i, \boldsymbol{\theta}^*]\mathbf{D}^{-1}\right) \\ &\quad + \frac{1}{\sigma^2}\text{tr}\left(E\left[u_i\boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i^\top|\mathbf{Q}_i, \mathbf{C}_i, \boldsymbol{\theta}^*\right]\boldsymbol{\Omega}_{n_i}^{-1}\right)\Big\}. \end{aligned}$$

The differential with respect to ρ, is

$$\begin{aligned} \frac{\partial Q(\boldsymbol{\theta}|\boldsymbol{\theta}^*)}{\partial\rho} &= -\frac{1}{2}\sum_{i=1}^{n}\left[\frac{\partial\log|\boldsymbol{\Omega}_{n_i}|}{\partial\rho} + \frac{1}{\sigma^2}\frac{\partial\text{tr}\left(E\left[u_i\boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i^\top|\mathbf{Q}_i, \mathbf{C}_i, \boldsymbol{\theta}^*\right]\boldsymbol{\Omega}_{n_i}^{-1}\right)}{\partial\rho}\right] \\ &= -\frac{1}{2}\sum_{i=1}^{n}\left[\text{tr}\left(\boldsymbol{\Omega}_{n_i}^{-1}\frac{\partial\boldsymbol{\Omega}_{n_i}}{\partial\rho}\right) - \frac{1}{\sigma^2}\text{tr}\left(\boldsymbol{\Omega}_{n_i}^{-1}\frac{\partial\boldsymbol{\Omega}_{n_i}^{-1}}{\partial\rho}\boldsymbol{\Omega}_{n_i}^{-1}E\left[u_i\boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i^\top|\mathbf{Q}_i, \mathbf{C}_i, \boldsymbol{\theta}^*\right]\right)\right] \\ &= -\frac{1}{2}\sum_{i=1}^{n}\text{tr}\left(\boldsymbol{\Omega}_{n_i}^{-1}\dot{\boldsymbol{\Omega}}_{n_i} - \frac{1}{\sigma^2}\boldsymbol{\Omega}_{n_i}^{-1}\dot{\boldsymbol{\Omega}}_{n_i}\boldsymbol{\Omega}_{n_i}^{-1}E\left[u_i\boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i^\top|\mathbf{Q}_i, \mathbf{C}_i, \boldsymbol{\theta}^*\right]\right) \\ &= -\frac{1}{2}\sum_{i=1}^{n}\text{tr}\left[\left(\boldsymbol{\Omega}_{n_i}^{-1} - \frac{1}{\sigma^2}\boldsymbol{\Omega}_{n_i}^{-1}E\left[u_i\boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i^\top|\mathbf{Q}_i, \mathbf{C}_i, \boldsymbol{\theta}^*\right]\boldsymbol{\Omega}_{n_i}^{-1}\right)\dot{\boldsymbol{\Omega}}_{n_i}\right], \end{aligned}$$

where $\dot{\boldsymbol{\Omega}}_{n_i} = \frac{\partial\boldsymbol{\Omega}_{n_i}}{\partial\rho}$ and

$$E\left[u_i\boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i^\top|\mathbf{Q}_i, \mathbf{C}_i, \boldsymbol{\theta}^*\right] = E\left[u_i(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{Z}_i\mathbf{b}_i)(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{Z}_i\mathbf{b}_i)^\top|\mathbf{Q}_i, \mathbf{C}_i, \boldsymbol{\theta}^*\right].$$

Therefore the parameter ρ is updated solving $\frac{\partial Q(\boldsymbol{\theta}|\boldsymbol{\theta}^*)}{\partial\rho} = 0$.

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

$$\mathbf{y}_i \overset{\text{ind.}}{\sim} Tt_{n_i}(\mathbf{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma}_i, \nu),$$

where $\boldsymbol{\Sigma}_i = \sigma^2\boldsymbol{\Omega}_{n_i} + \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^\top$

$$E(u_i|\mathbf{y}_i) = \frac{\nu + n_i}{\nu + \delta},$$

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

$$\mathbf{b}_i|\mathbf{y}_i, u_i \overset{\text{ind.}}{\sim} N_q\left(\frac{u_i}{\sigma^2}\left(u_i\mathbf{D}^{-1} + \frac{u_i}{\sigma^2}\mathbf{Z}_i^\top\boldsymbol{\Omega}_{n_i}^{-1}\mathbf{Z}_i\right)^{-1}\mathbf{Z}_i^\top\boldsymbol{\Omega}_{n_i}^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}), \left(u_i\mathbf{D}^{-1} + \frac{u_i}{\sigma^2}\mathbf{Z}_i^\top\boldsymbol{\Omega}_{n_i}^{-1}\mathbf{Z}_i\right)^{-1}\right),$$

$$\mathbf{b}_i|\mathbf{y}_i, u_i \overset{\text{ind.}}{\sim} N_q\left(\boldsymbol{\varphi}_i(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}), \frac{\sigma^2}{u_i}\boldsymbol{\Lambda}_i\right),$$

with $\boldsymbol{\Lambda}_i = (\sigma^2\mathbf{D}^{-1} + \mathbf{Z}_i^\top\boldsymbol{\Omega}_{n_i}^{-1}\mathbf{Z}_i)^{-1}$ and $\boldsymbol{\varphi}_i = \boldsymbol{\Lambda}_i\mathbf{Z}_i^\top\boldsymbol{\Omega}_{n_i}^{-1}$.

Like as in Appendix A, the expectation term are:

$$\widehat{u\mathbf{y}_i} = E\{u_i\mathbf{y}_i|\mathbf{Q}_i, \mathbf{C}_i, \boldsymbol{\theta}^*\} = \frac{T_{n_i}(a|\boldsymbol{\mu}, \boldsymbol{\Sigma}^*, \nu+2)}{T_{n_i}(a|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)}E\{\mathbf{W}_i\},$$

$$\widehat{u\mathbf{y}_i^2} = E\{u_i\mathbf{y}_i\mathbf{y}_i^\top|\mathbf{Q}_i, \mathbf{C}_i, \boldsymbol{\theta}^*\} = \frac{T_{n_i}(a|\boldsymbol{\mu}, \boldsymbol{\Sigma}^*, \nu+2)}{T_{n_i}(a|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)}E\{\mathbf{W}_i\mathbf{W}_i^\top\},$$

$$\widehat{u_i} = E\{u_i|\mathbf{Q}_i, \mathbf{C}_i, \boldsymbol{\theta}^*\} = \frac{T_{n_i}(a|\boldsymbol{\mu}, \boldsymbol{\Sigma}^*, \nu+2)}{T_{n_i}(a|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)},$$

$$\widehat{u\mathbf{b}_i} = E\{u_i\mathbf{b}_i|\mathbf{Q}_i, \mathbf{C}_i, \boldsymbol{\theta}^*\} = \boldsymbol{\varphi}_i\left[\widehat{u\mathbf{y}_i} - \mathbf{X}_i\boldsymbol{\beta}\widehat{u_i}\right],$$

$$\begin{aligned}\widehat{u\mathbf{b}_i^2} &= E\{u_i\mathbf{b}_i\mathbf{b}_i^\top|\mathbf{Q}_i, \mathbf{C}_i, \boldsymbol{\theta}^*\} \\ &= \boldsymbol{\Lambda}_i\sigma^2 + \boldsymbol{\varphi}_i\left(\widehat{u\mathbf{y}_i^2} - \widehat{u\mathbf{y}_i}\boldsymbol{\beta}^\top\mathbf{X}_i^\top - \mathbf{X}_i\boldsymbol{\beta}\widehat{u\mathbf{y}_i}^\top + \widehat{u_i}\mathbf{X}_i\boldsymbol{\beta}\boldsymbol{\beta}^\top\mathbf{X}_i^\top\right)\boldsymbol{\varphi}_i^\top,\end{aligned}$$

$$\widehat{u\mathbf{y}\mathbf{b}_i} = E\{u_i\mathbf{b}_i\mathbf{b}_i^\top|\mathbf{Q}_i, \mathbf{C}_i, \boldsymbol{\theta}^*\} = \left[\widehat{u\mathbf{y}_i^2} - \widehat{u\mathbf{y}_i}\boldsymbol{\beta}^\top\mathbf{X}_i^\top\right]\boldsymbol{\varphi}_i^\top,$$

So the value of $E\left[u_i\boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i^\top|\mathbf{Q}_i, \mathbf{C}_i, \boldsymbol{\theta}^*\right]$ is given by

$$\begin{aligned}E\left[u_i\boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_i^\top|\mathbf{Q}_i, \mathbf{C}_i, \boldsymbol{\theta}^*\right] &= E\left[u_i(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{Z}_i\mathbf{b}_i)(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{Z}_i\mathbf{b}_i)^\top|\mathbf{Q}_i, \mathbf{C}_i, \boldsymbol{\theta}^*\right] \\ &= \operatorname{tr}(\widehat{u\mathbf{y}_i^2}) - \widehat{u\mathbf{y}_i}^\top\mathbf{X}_i\boldsymbol{\beta} - \operatorname{tr}(\widehat{u\mathbf{y}\mathbf{b}_i}^\top\mathbf{Z}_i) - \boldsymbol{\beta}^\top\mathbf{X}_i^\top\widehat{u\mathbf{y}_i} + \boldsymbol{\beta}^\top\mathbf{X}_i^\top\widehat{u_i}\mathbf{X}_i\boldsymbol{\beta} \\ &\quad + \boldsymbol{\beta}^\top\mathbf{X}_i^\top\mathbf{Z}_i\widehat{u\mathbf{b}_i} - \operatorname{tr}(\widehat{u\mathbf{y}\mathbf{b}_i}\mathbf{Z}_i^\top) + \widehat{u\mathbf{b}_i}^\top\mathbf{Z}_i^\top\mathbf{X}_i\boldsymbol{\beta} + \operatorname{tr}(\widehat{u\mathbf{b}_i^2}\mathbf{Z}_i^\top\mathbf{Z}_i).\end{aligned}$$

References

Arellano-Valle, R. and M. Genton (2010). Multivariate extended skew-t distributions and related families. *Metron LXVIII*, 201–234.

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

Genz, A., F. Bretz, T. Hothorn, T. Miwa, X. Mi, F. Leisch, and F. Scheipl (2008). mvtnorm:

Multivariate Normal and t Distribution. *R package version 0.9-2, URL http://CRAN. R-project. org/package= mvtnorm*.

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

Lachos, V., D. Bandyopadhyay, and D. Dey (2011). Linear and nonlinear mixed-effects models for censored hiv viral loads using normal/independent distributions. *Biometrics 67*, 1594–1604.

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

Lin, T., H. Ho, H. Chen, and W. Wang (2011). Some results on the truncated multivariate t distribution. *Journal of Statistical Planning and Inference*.

Lin, T. and J. Lee (2007). Bayesian analysis of hierarchical linear mixed modeling using the multivariate t distribution. *Journal of Statistical Planning and Inference 137*(2), 484–495.

Lindstrom, M. and D. Bates (1990). Nonlinear mixed-effects models for repeated-measures data. *Biometrics 46*, 673687.

Liu, C. (1996). Bayesian robust multivariate linear regression with incomplete data. *Journal of the American Statistical Association 91*, 1219–1227.

Louis, T. (1982). Finding the observed information matrix when using the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 226–233.

Lucas, A. (1997). Robustness of the student t based M-estimator. *Communications in Statistics-Theory and Methods 26*(5), 1165–1182.

Matos, L. A., V. Lachos, N. Balakrishnan, and F. Labra (2011). Influence diagnostics in linear and nonlinear mixed-effects models with censored data. *Universidade Estadual de Campinas Technical Report, 10-11*, 484–496.

McLachlan, G. and T. Krishnan (1996). *The EM Algorithm and Extensions.* Wiley.

Meng, X. and D. B. Rubin (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika 81*, 633–648.

Meza, C., F. Osorio, and R. De la Cruz (2011). Estimation in nonlinear mixed-effects models using heavy-tailed distributions. *Statistics and Computing*.

Pinheiro, J. C. and M. Bates, Douglas (2000). *Mixed-Effects Models in S and S-PLUS*. New York, NY: Springer.

Pinheiro, J. C., C. H. Liu, and Y. N. Wu (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 Development Core Team (2009). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Rosa, G. J. M., C. R. Padovani, and D. Gianola (2003). Robust linear mixed models with normal/independent distributions and bayesian mcmc implementation. *Biometrical Journal 45*, 573–590.

Saitoh, A., M. Foca, R. Viani, S. Heffernan-Vacca, F. Vaida, J. Lujan-Zilbermann, P. Emmanuel, J. Deville, and S. Spector (2008). Clinical outcomes after an unstructured treatment interruption in children and adolescents with perinatally acquired HIV infection. *Pediatrics 121*(3), e513–e521.

Self, S. and K. Liang (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, 605–610.

Song, P., P. Zhang, and A. Qu (2007). Maximum likelihood inference in robust linear mixed-effects models using multivariate t distributions. *Statistica Sinica 17*, 929–943.

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

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

Wang, W. and T. Fan (2011). Estimation in multivariate t linear mixed models for multivariate longitudinal data. *Statistica Sinica 21*, 1857–1880.

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