Quantile Regression for Nonlinear Mixed Effects Models: A Likelihood Based Perspective

Christian E. Galarza[a] Luis M. Castro[b] Francisco Louzada[c]
Victor H. Lachos[a]*

[a]*Departamento de Estatística, Universidade Estadual de Campinas, Campinas, Brazil*
[b]*Departamento de Estadística and CI^{2}MA, Universidad de Concepción, Chile*
[c]*Department of Applied Mathematics and Statistics, Universidade de São Paulo, São Carlos, Brazil.*

Abstract

Longitudinal data are frequently analyzed using normal mixed effects models. Moreover, the traditional estimation methods are based on mean regression, which leads to non-robust parameter estimation for non-normal error distributions. Compared to the conventional mean regression approach, quantile regression (QR) can characterize the entire conditional distribution of the outcome variable and is more robust to the presence of outliers and misspecification of the error distribution. This paper develops a likelihood-based approach to analyzing QR models for correlated continuous longitudinal data via the asymmetric Laplace (AL) distribution. Exploiting the nice hierarchical representation of the AL distribution, our classical approach follows the Stochastic Approximation of the EM (SAEM) algorithm for deriving exact maximum likelihood estimates of the fixed-effects and variance components in nonlinear mixed effects models (NLMEMs). We evaluate the finite sample performance of the algorithm and the asymptotic properties of the ML estimates through empirical experiments and applications to two real life datasets. The proposed SAEM algorithm is implemented in the `R` package `qrNLMM`.

Keywords Asymmetric Laplace distribution; Nonlinear mixed effects models; Quantile regression; SAEM algorithm; Stochastic Approximations.

1 Introduction

Non-linear mixed-effects (NLME) models are frequently used for analyzing grouped data, clustered data, longitudinal data, multilevel data, among others. This is because, this type of models allows us to deal with non-linear relationships between the observed response and the covariates and/or random effects, and at the same time, takes into account within and between-subject correlations in the statistical modelling of the observed data. In general, NLME models arise as a

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

consequence of the mathematical modelling of biological, chemical and physics phenomena, using known families of non-linear functions with attractive properties such as the asymptotic, uniqueness of maximum value, monotonicity and positive range (Pinheiro & Bates, 2000; Davidian & Giltinan, 2003; Wu, 2010). Although most of the currently NLME models research are focused on the estimation of the conditional mean of the response given some covariates, sometimes the estimation of this quantity presents a lack of meaning, specially when the conditional distribution of the response (given the covariates) is asymmetric, multimodal, or simply, is severely affected by atypical observations (outliers). In this case, conditional quantile regression (QR) methods (Koenker, 2004, 2005) become in a more appropriate strategy for describing the conditional distribution of the outcome variable given the covariates. One of the advantages for using QR methods is that it does not impose any distribution assumption on the error term, except that this term must have a conditional quantile equal to zero. Moreover, and from a practical point of view, standard QR methods are already implemented in statistical software such `R` in its package `quantreg()`.

QR methods was initially developed under a univariate framework but, nowadays, the abundance of correlated data in real-life applications have generated the study of several extensions of QR methods based on mixed models. Some of these extensions considers the distribution-free approach (Lipsitz *et al.*, 1997; Galvao & Montes-Rojas, 2010; Galvao Jr, 2011; Fu & Wang, 2012), and others consider the traditional likelihood-based approach using the asymmetric Laplace (AL) distribution (Geraci & Bottai, 2007; Yuan & Yin, 2010; Geraci & Bottai, 2014). In this context, Geraci & Bottai (2007) proposed a Monte Carlo EM (MCEM) algorithm for the QR model considering continuous responses with a subject-specific univariate random intercept. Recently, Geraci & Bottai (2014) extended their previous work by considering a general linear quantile mixed effects regression (QR-LME) model with multiple random effects. Note that, in that work, the authors considered the estimation of the fixed effects and the covariance components through an efficient combination of Gaussian quadrature approximations and non-smooth optimisation algorithms. On the other hand, Yuan & Yin (2010) extend the QR model proposed by Geraci & Bottai (2007) to case of linear mixed effects models for longitudinal measurements with missing data. From the non-linear point of view, Wang (2012) considered a QR-NLME model from a Bayesian perspective, showing that this model may be a better alternative than the mean regression estimation under the presence of asymmetric and multimodal data.

Although some results based on QR-NLME models have recently appeared in the statistical literature, to the best of our knowledge, there seem to be no studies and contributions considering an exact inference for QR-NLME models from a likelihood-based perspective. For that reason, the aim of our paper is to propose a QR-NLME model model using the AL distribution and considering a full likelihood-based inference via the implementation of the stochastic version of the EM (SAEM) algorithm, proposed by Delyon *et al.* (1999) for the maximum likelihood (ML) estimation. The SAEM algorithm has been proved to be a more computationally efficient algorithm than the classical MCEM due to the recycling of simulations from one iteration to the next in the smoothing phase of it. Moreover, as pointed out by Meza *et al.* (2012), the SAEM algorithm, unlike the MCEM, converges even in a typically small simulation size. Recently, Kuhn & Lavielle (2005) showed that the SAEM algorithm is very efficient for computing the ML estimates in mixed effects models. It is important to stress that, the empirical results shows that the ML estimates based on our proposed SAEM algorithm provide good asymptotic properties. Moreover, the application of

our method is conducted using the recently R package so-called qrNLMM().

The rest of the paper proceeds as follows. Section 2 presents some preliminaries results, particularly, the connection between QR models and and AL distribution. In this Section, an outline of the EM and SAEM algorithms are also presented. Section 3 provides the MCEM and SAEM algorithms for a general NLME model, while Section 4 outlines the likelihood-based estimation and standard errors of the parameter estimates under the proposed model. Section 5 presents some simulation studies. The analysis of two longitudinal datasets are presented in Section 6. Finally, Section 7 concludes the paper, sketching our plan of future work.

2 Preliminaries

In this section, we provide some useful results related to the AL distribution and QR model. We also present some background about the EM and SAEM algorithms for the ML estimation.

2.1 Connection between QR model and AL distribution

Let $\mathbf{y}_i$ denote the response of interest and $\mathbf{x}_i$ the corresponding covariate vector of dimension $k \times 1$ for subject i, $i = 1, \ldots, n$. Then, the pth $(0 < p < 1)$ QR model takes the form

$$Q_p(\mathbf{y}_i) = \mathbf{x}_i^\top \boldsymbol{\beta}_p, \ i = 1, \ldots, n,$$

where $Q_p(\mathbf{y}_i)$ is the quantile function (or the inverse cumulative distribution function) of $\mathbf{y}_i$ given $\mathbf{x}_i$ evaluated at p, and $\boldsymbol{\beta}_p$ is a vector of regression parameters corresponding to the pth quantile. The regression vector $\boldsymbol{\beta}_p$ is estimated by minimizing

$$\sum_{i=1}^{n} \rho_p(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}_p), \tag{1}$$

where $\rho_p(\cdot)$ is the check (or loss) function defined by $\rho_p(u) = u(p - \mathbb{I}\{u < 0\})$, with $\mathbb{I}\{\cdot\}$ the usual indicator function.

Next, we define the AL distribution. A random variable Y is distributed as an AL distribution Yu & Moyeed (2001) with location parameter μ, scale parameter $\sigma > 0$ and skewness parameter $p \in (0,1)$, if its probability density function (pdf) given by

$$f(y|\mu, \sigma, p) = \frac{p(1-p)}{\sigma} \exp\left\{-\rho_p\left(\frac{y-\mu}{\sigma}\right)\right\}. \tag{2}$$

The AL distribution is an asymmetric distribution with a straightforward skewness parametrization, and the check function $\rho_p(\cdot)$ is closely related to the AL distribution Koenker & Machado (1999); Yu & Moyeed (2001). Note that minimizing the loss function in (1) is equivalent to maximizing the AL distribution likelihood function. This is in tune to the result from simple linear regression, where the ordinary least square (OLS) estimator of the regression parameter minimizing the error sum of squares is equivalent to the maximum likelihood (ML) estimator of the corresponding Gaussian likelihood. Note that $W = \rho_p\left(\frac{Y-\mu}{\sigma}\right)$ follows an exponential distribution with parameter

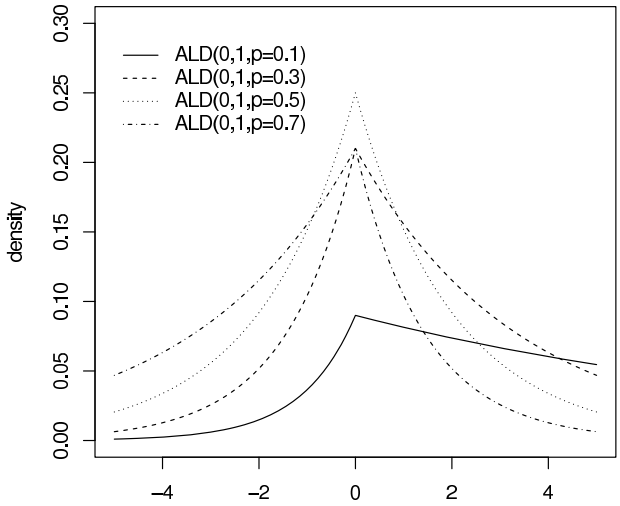

Figure 1: Standard asymmetric Laplace density function

equal to 1. Figure 1 plots the AL distribution for different values of p. For example, when $p = 0.1$, most of the mass is concentrated around the right tail, while for $p = 0.5$, both tails of the distribution have equal mass.

The AL distribution has a useful stochastic representation (Kotz *et al.*, 2001; Kuzobowski & Podgorski, 2000). Let $U \sim \exp(\sigma)$ and $Z \sim \mathrm{N}(0,1)$ be two independent random variables. Then, $Y \sim \mathrm{AL}(\mu, \sigma, p)$ can be represented as

$$Y \stackrel{d}{=} \mu + \vartheta_p U + \tau_p \sqrt{\sigma U} Z, \tag{3}$$

where $\vartheta_p = \frac{1-2p}{p(1-p)}$, $\tau_p^2 = \frac{2}{p(1-p)}$ and $\stackrel{d}{=}$ denotes equality in distribution. This representation is useful for obtaining the moment generating function (*mgf*) and implementing the EM algorithm. From (3), the hierarchical representation of the AL distribution is given by

$$\begin{aligned} Y \mid U = u &\sim \mathrm{N}(\mu + \vartheta_p u, \tau_p^2 \sigma u), \\ U &\sim \exp(\sigma). \end{aligned} \tag{4}$$

Moreover, $U \mid Y \sim \mathrm{GIG}(\frac{1}{2}, \delta, \gamma)$, where $\mathrm{GIG}(\nu, a, b)$ represents the Generalized Inverse Gaussian (GIG) distribution (Barndorff-Nielsen & Shephard, 2001) with *pdf* given by

$$h(u \mid \nu, a, b) = \frac{(b/a)^\nu}{2K_\nu(ab)} u^{\nu-1} \exp\Big\{ -\frac{1}{2}\big(a^2/u + b^2 u\big) \Big\}, \quad u > 0, \quad , \nu \in \mathbb{R}, \quad a, b > 0,$$

where $K_\nu(\cdot)$ denotes the modified Bessel function of the third kind. The moments of U can be expressed as

$$E[U^k] = \left(\frac{a}{b}\right)^k \frac{K_{\nu+k}(ab)}{K_\nu(ab)}, \quad k \in \mathbb{R}. \tag{5}$$

2.2 The EM and SAEM algorithms

In models with missing data, the EM algorithm (Dempster *et al.*, 1977) has established itself as the centerpiece for ML estimation of model parameters, mostly when the maximization of the

observed log-likelihood function denoted by $\ell(\boldsymbol{\theta}; \mathbf{y}_{\text{obs}}) = \log f(\mathbf{y}_{\text{obs}}; \boldsymbol{\theta})$ is complicated. Let y_{obs} and $\mathbf{q}$ represent observed and missing data, respectively, such that the complete data can be written as $\mathbf{y}_{\text{com}} = (\mathbf{y}_{\text{obs}}, \mathbf{q})^\top$. This iterative algorithm maximizes the complete log-likelihood function $\ell_c(\boldsymbol{\theta}; \mathbf{y}_{\text{com}}) = \log f(\mathbf{y}_{\text{obs}}, \mathbf{q}; \boldsymbol{\theta})$ at each step, converging to a stationary point of the observed likelihood $\ell(\boldsymbol{\theta}; \mathbf{y}_{\text{obs}})$ under mild regularity conditions (Wu, 1983; Vaida, 2005). The EM algorithm proceeds in two simple steps:

E-Step: Replace the observed likelihood by the complete likelihood and compute its conditional expectation $Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k)}) = \mathrm{E}\{\ell_c(\boldsymbol{\theta}; \mathbf{y}_{\text{com}})|\widehat{\boldsymbol{\theta}}^{(k)}, \mathbf{y}_{\text{obs}}\}$, where $\widehat{\boldsymbol{\theta}}^{(k)}$ is the estimate of $\boldsymbol{\theta}$ at the k-th iteration;
M-Step: Maximize $Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k)})$ with respect to $\boldsymbol{\theta}$ to obtain $\widehat{\boldsymbol{\theta}}^{(k+1)}$.

However, in some situations, the E-step cannot be obtained analytically and has to be calculated though a simulation step. Wei & Tanner (1990) proposed the Monte Carlo EM (MCEM) algorithm in which the E-step is replaced by a Monte Carlo approximation based on a large number of independent simulations of the missing data/latent variables. This simple solution is in fact computationally expensive because the large number of independent simulations of the missing data/latent variables required to achieve a good approximation of the algorithm. Consequently, in order to reduce the amount of simulations required by the MCEM algorithm, the SAEM algorithm proposed by Delyon *et al.* (1999) replaces the E-step by a stochastic approximation procedure. Besides having good theoretical properties, the SAEM algorithm estimates the population parameters accurately, converging to the global maxima of the ML estimates under quite general conditions (Allassonnière *et al.*, 2010; Delyon *et al.*, 1999; Kuhn & Lavielle, 2004).

At each iteration, the SAEM algorithm successively simulates missing data/latent variables using their conditional distributions, updating the model parameters. Thus, at iteration k, the SAEM algorithm proceeds as follows:

E-Step:

- Simulation: Draw $(\mathbf{q}^{(\ell,k)})$, $\ell = 1, \ldots, m$ from the conditional distribution of the missing data $f(\mathbf{q}|\boldsymbol{\theta}^{(k-1)}, \mathbf{y}_{\text{obs}})$.

- Stochastic Approximation: Update the $Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k)})$ function as

$$Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k)}) \approx Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k-1)}) + \delta_k \left[\frac{1}{m} \sum_{\ell=1}^{m} \log f(\mathbf{y}_{\text{obs}}, \mathbf{q}^{(\ell,k)}; \boldsymbol{\theta}) - Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k-1)}) \right] \tag{6}$$

M-Step:

- Maximization: Update $\widehat{\boldsymbol{\theta}}^{(k)}$ as $\widehat{\boldsymbol{\theta}}^{(k+1)} = \arg\max_{\theta} Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k)})$,

Note that, although the E-Step is similar in the SAEM and MCEM algorithms, a small number of simulations m (for practical situations, $m \leq 20$ is suggested) is necessary in the first one. This is possible because, unlike the traditional EM algorithm and its variants, the SAEM algorithm uses

not only the current simulation of the missing data/latent variables at the iteration k but some or all the previous simulations. In fact, this 'memory' property is set by the smoothing parameter δ_k. In our case, we suggested the following choice of the smoothing parameter given as

$$\delta_k = \begin{cases} 1, & \text{for} \quad 1 \leq k \leq cW \\ \frac{1}{k-cW}, & \text{for} \quad cW+1 \leq k \leq W \end{cases}$$

where W is the maximum number of iterations, and c a cut point ($0 \leq c \leq 1$) which determines the percentage of initial iterations with no memory.

3 The QR non-linear mixed model

3.1 The model

In this Section, we proposed the following general mixed-effects model. Let $\mathbf{y}_i = (y_{i1}, \ldots, y_{in_i})^\top$ be the continuous response for subject i and $\eta = (\eta(\boldsymbol{\phi}_i, x_{i1}), \ldots, \eta(\boldsymbol{\phi}_i, x_{in_i}))^\top$ a nonlinear differentiable function of vector-valued random parameters ϕ_i of dimension r. Moreover, let $\mathbf{x}_i$ be a matrix of covariates of dimension $n_i \times r$. The NLME model is defined as

$$\mathbf{y}_i = \eta(\boldsymbol{\phi}_i, \mathbf{x}_i) + \boldsymbol{\varepsilon}_i, \qquad \boldsymbol{\phi}_i = \mathbf{A}_i \boldsymbol{\beta}_p + \mathbf{B}_i \mathbf{b}_i, \tag{7}$$

where $\mathbf{A}_i$ and $\mathbf{B}_i$ are design matrices (fixed) of dimensions $r \times d$ and $r \times q$, respectively, possibly depending on elements of $\mathbf{x}_i$ and incorporating time varying covariates in fixed or random effects, $\boldsymbol{\beta}_p$ is the regression coefficient corresponding to the pth quantile, $\mathbf{b}_i$ is a q-dimensional random effects vector associated to the i-th subject and $\boldsymbol{\varepsilon}_i$ the independent and identically distributed vector of random errors. We define the pth quantile function of the response y_{ij} as

$$Q_p(y_{ij} \mid x_{ij}, \mathbf{b}_i) = \eta(\boldsymbol{\phi}_i, x_{ij}) = \eta(\mathbf{A}_i \boldsymbol{\beta}_p + \mathbf{B}_i \mathbf{b}_i, x_{ij}). \tag{8}$$

where Q_p denotes the inverse of the unknown distribution function F. In this setting, the random effects $\mathbf{b}_i$ are distributed independent and identically distributed (*i.i.d*) as $\mathrm{N}_q(\mathbf{0}, \boldsymbol{\Psi})$, where the dispersion matrix $\boldsymbol{\Psi} = \boldsymbol{\Psi}(\boldsymbol{\alpha})$ depends on unknown and reduced parameters $\boldsymbol{\alpha}$. The error terms are distributed as $\varepsilon_{ij} \overset{\text{iid}}{\sim} \mathrm{AL}(0, \sigma)$, being uncorrelated with the random effects. Then, conditionally on $\mathbf{b}_i$, the observed responses for subject i, *i.e.*, y_{ij} for $j = 1, \ldots, n_i$ are independent following an AL distribution with *pdf* given by

$$f(y_{ij} \mid \boldsymbol{\beta}_p, \mathbf{b}_i, \sigma) = \frac{p(1-p)}{\sigma} \exp\left\{ -\rho_p \left(\frac{y_{ij} - \eta(\mathbf{A}_i \boldsymbol{\beta}_p + \mathbf{B}_i \mathbf{b}_i, \mathbf{x}_{ij})}{\sigma} \right) \right\}. \tag{9}$$

3.2 The MCEM algorithm

In this section we develop a MCEM algorithm for the ML estimation of the parameters in the QR-NLME model. This model has a flexible hierarchical representation, which is useful for deriving

interesting theoretical properties. From (4), we have that the QR-NLME model defined in (8)-(9), can be represented as follows:

$$\begin{aligned} \mathbf{y}_i \mid \mathbf{b}_i, \mathbf{u}_i &\sim \mathrm{N}_{n_i}\left(\boldsymbol{\eta}(\mathbf{A}_i\boldsymbol{\beta}_p + \mathbf{B}_i\mathbf{b}_i, \mathbf{x}_i) + \vartheta_p\mathbf{u}_i, \sigma\tau_p^2\boldsymbol{D}_i\right), \\ \mathbf{b}_i &\sim \mathrm{N}_q\left(\mathbf{0}, \boldsymbol{\Psi}\right), \\ \mathbf{u}_i &\sim \prod_{j=1}^{n_i}\exp(\sigma), \end{aligned} \tag{10}$$

for $i = 1, \ldots, n$, where ϑ_p and τ_p^2 are given as in (3); $\boldsymbol{D}_i$ represents a diagonal matrix that contains the vector of latent variables $\mathbf{u}_i = (u_{i1}, \ldots, u_{in_i})^\top$ and $\exp(\sigma)$ denotes the exponential distribution with mean σ. Let $\mathbf{y}_{ic} = (\mathbf{y}_i^\top, \mathbf{b}_i^\top, \mathbf{u}_i^\top)^\top$, with $\mathbf{y}_i = (y_{i1}, \ldots, y_{in_i})^\top$, $\mathbf{b}_i = \left(b_{i1}, \ldots, b_{iq}\right)^\top$, $\mathbf{u}_i = (u_{i1}, \ldots, u_{in_i})^\top$ and let $\boldsymbol{\theta}^{(k)} = (\boldsymbol{\beta}_p^{(k)\top}, \sigma^{(k)}, \boldsymbol{\alpha}^{(k)\top})^\top$, the estimate of $\boldsymbol{\theta}$ at the k-th iteration. Since $\mathbf{b}_i$ and $\mathbf{u}_i$ are independent for all $i = 1, \ldots, n$, it follows from (4) that the complete-data log-likelihood function is given by

$$\ell_c(\boldsymbol{\theta}; \mathbf{y}_c) = \sum_{i=1}^{n}\ell_c(\boldsymbol{\theta}; \mathbf{y}_{ic}),$$

where

$$\begin{aligned} \ell_c(\boldsymbol{\theta}; \mathbf{y}_{ic}) &= \text{constant} - \frac{3}{2}n_i log\sigma - \frac{1}{2}log\left|\boldsymbol{\Psi}\right| - \frac{1}{2}\mathbf{b}_i^\top\boldsymbol{\Psi}^{-1}\mathbf{b}_i - \frac{1}{\sigma}\mathbf{u}_i^\top\mathbf{1}_{n_i} \\ &\quad - \frac{1}{2\sigma\tau_p^2}\left(\mathbf{y}_i - \boldsymbol{\eta}(\mathbf{A}_i\boldsymbol{\beta}_p + \mathbf{B}_i\mathbf{b}_i, \mathbf{x}_i) - \vartheta_p\mathbf{u}_i\right)^\top\boldsymbol{D}_i^{-1}\left(\mathbf{y}_i - \boldsymbol{\eta}(\mathbf{A}_i\boldsymbol{\beta}_p + \mathbf{B}_i\mathbf{b}_i, \mathbf{x}_i) - \vartheta_p\mathbf{u}_i\right), \end{aligned} \tag{11}$$

where $\mathbf{1}_p$ is a vector of ones of dimension p. Since $\mathbf{A}_i$, $\mathbf{B}_i$ and $\mathbf{x}_i$ are known matrices, we simplify the notation by writing $\boldsymbol{\eta}(\boldsymbol{\beta}_p, \mathbf{b}_i)$ to represent $\boldsymbol{\eta}(\boldsymbol{\phi}_i, \mathbf{x}_i) = \boldsymbol{\eta}(\mathbf{A}_i\boldsymbol{\beta}_p + \mathbf{B}_i\mathbf{b}_i, \mathbf{x}_i)$. Given the current estimate $\boldsymbol{\theta} = \boldsymbol{\theta}^{(k)}$, the E-step calculates the function

$$Q\left(\boldsymbol{\theta} \mid \widehat{\boldsymbol{\theta}}^{(k)}\right) = \sum_{i=1}^{n}Q_i\left(\boldsymbol{\theta} \mid \widehat{\boldsymbol{\theta}}^{(k)}\right),$$

where

$$\begin{aligned} Q_i\left(\boldsymbol{\theta} \mid \widehat{\boldsymbol{\theta}}^{(k)}\right) &= \mathrm{E}\left\{\ell_c(\boldsymbol{\theta}; \mathbf{y}_{ic}) \mid \boldsymbol{\theta}^{(k)}, \mathbf{y}\right\} \\ &\propto -\frac{3}{2}n_i log\sigma - \frac{1}{2}log\left|\boldsymbol{\Psi}\right| - \frac{1}{2}\mathrm{tr}\left\{\widehat{(\mathbf{b}\mathbf{b}^\top)_i}^{(k)}\boldsymbol{\Psi}^{-1}\right\} - \frac{1}{2\sigma\tau_p^2}\left[\mathbf{y}_i^\top\widehat{\boldsymbol{D}_i^{-1}}^{(k)}\mathbf{y}_i\right. \\ &\quad \left. - 2\vartheta_p\mathbf{y}_i^\top\mathbf{1}_{n_i} + \frac{\tau_p^4}{4}\widehat{\mathbf{u}}_i^{(k)\top}\mathbf{1}_{n_i} - 2\mathbf{y}_i^\top(\widehat{\boldsymbol{D}^{-1}\boldsymbol{\eta}})_i^{(k)} + 2\vartheta_p\mathbf{1}_{n_i}^\top\widehat{\boldsymbol{\eta}}_i^{(k)} + \widehat{\boldsymbol{\eta}_i^\top\boldsymbol{D}_i^{-1}\boldsymbol{\eta}_i}^{(k)}\right], \end{aligned} \tag{12}$$

where $\boldsymbol{\eta}_i = \boldsymbol{\eta}(\boldsymbol{\beta}_p, \mathbf{b}_i)$, $\mathrm{tr}(\mathbf{A})$ denotes the trace of matrix $\mathbf{A}$. The calculation of this function requires the following expressions

$$\widehat{\boldsymbol{\eta}}_i^{(k)} = \mathrm{E}\left\{\boldsymbol{\eta}_i \mid \boldsymbol{\theta}^{(k)}, \mathbf{y}_i\right\}, \qquad \widehat{\mathbf{u}}_i^{(k)} = \mathrm{E}\left\{\mathbf{u}_i \mid \boldsymbol{\theta}^{(k)}, \mathbf{y}_i\right\},$$

$$\widehat{(\mathbf{b}\mathbf{b}^\top)_i}^{(k)} = \mathrm{E}\left\{\mathbf{b}_i\mathbf{b}_i^\top \mid \boldsymbol{\theta}^{(k)}, \mathbf{y}_i\right\}, \qquad \widehat{\boldsymbol{D}_i^{-1}}^{(k)} = \mathrm{E}\left\{\boldsymbol{D}_i^{-1} \mid \boldsymbol{\theta}^{(k)}, \mathbf{y}_i\right\},$$

$$(\widehat{\boldsymbol{D}^{-1}\boldsymbol{\eta}})_i^{(k)} = \mathrm{E}\left\{\boldsymbol{D}_i^{-1}\boldsymbol{\eta}_i^{(k)} \mid \boldsymbol{\theta}^{(k)}, \mathbf{y}_i\right\}, \qquad (\widehat{\boldsymbol{\eta}^\top\boldsymbol{D}^{-1}\boldsymbol{\eta}})_i^{(k)} = \mathrm{E}\left\{\boldsymbol{\eta}_i^\top\boldsymbol{D}_i^{-1}\boldsymbol{\eta}_i \mid \boldsymbol{\theta}^{(k)}, \mathbf{y}_i\right\},$$

which do not have closed forms. Since the joint distribution of the latent variables $\left(\mathbf{b}_i^{(k)}, \mathbf{u}_i^{(k)}\right)$ is unknown and the conditional expectations cannot be computed analytically, for any function $g(\cdot)$, the MCEM algorithm approximates these expectations using a Monte Carlo approximation given by

$$\mathrm{E}[g(\mathbf{b}_i, \mathbf{u}_i) \mid \boldsymbol{\theta}^{(k)}, \mathbf{y}_i] \approx \frac{1}{m} \sum_{\ell=1}^{m} g(\mathbf{b}_i^{(\ell,k)}, \mathbf{u}_i^{(\ell,k)}), \tag{13}$$

which depend of the simulations of the two latent variables $\mathbf{b}_i^{(k)}$ and $\mathbf{u}_i^{(k)}$ from the conditional density $f(\mathbf{b}_i, \mathbf{u}_i \mid \boldsymbol{\theta}^{(k)}, \mathbf{y}_i)$. Using properties of conditional expectations, the expected value given in (13) can be more accurately approximated as

$$\begin{aligned} \mathrm{E}_{\mathbf{b}_i,\mathbf{u}_i}[g(\mathbf{b}_i, \mathbf{u}_i) \mid \boldsymbol{\theta}^{(k)}, \mathbf{y}_i] &= \mathrm{E}_{\mathbf{b}_i}[\mathrm{E}_{\mathbf{u}_i}[g(\mathbf{b}_i, \mathbf{u}_i) \mid \boldsymbol{\theta}^{(k)}, \mathbf{b}_i, \mathbf{y}_i] \mid \mathbf{y}_i] \\ &\approx \frac{1}{m} \sum_{\ell=1}^{m} \mathrm{E}_{\mathbf{u}_i}[g(\mathbf{b}_i^{(\ell,k)}, \mathbf{u}_i) \mid \boldsymbol{\theta}^{(k)}, \mathbf{b}_i^{(\ell,k)}, \mathbf{y}_i], \end{aligned} \tag{14}$$

where $\mathbf{b}^{(\ell,k)}$ is generated from $f(\mathbf{b}_i \mid \boldsymbol{\theta}^{(k)}, \mathbf{y}_i)$. Note that (14) is a more accurate approximation once it only depends of one MC approximation, instead of two (as is needed in (13)).

For generating random samples from the full conditional distribution $f(\mathbf{u}_i \mid \mathbf{y}_i, \mathbf{b}_i)$, first note that the vector $\mathbf{u}_i \mid \mathbf{y}_i, \mathbf{b}_i$ can be written as $\mathbf{u}_i \mid \mathbf{y}_i, \mathbf{b}_i = [\ u_{i1} \mid y_{i1}, \mathbf{b}_i,\ \ u_{i2} \mid y_{i2}, \mathbf{b}_i,\ \ \cdots\ \ , u_{in_i} \mid y_{in_i}, \mathbf{b}_i\ ]^\top$, since $u_{ij} \mid y_{ij}, \mathbf{b}_i$ is independent of $u_{ik} \mid y_{ik}, \mathbf{b}_i$, for all $j, k = 1, 2, \ldots, n_i$ and $j \neq k$. Thus, the distribution of $f(u_{ij} \mid y_{ij}, \mathbf{b}_i)$ is proportional to

$$f(u_{ij} \mid y_{ij}, \mathbf{b}_i) \propto \phi(y_{ij} \big| \eta_{ij}(\boldsymbol{\beta}_p, \mathbf{b}_i) + \vartheta_p u_{ij},\, \sigma\tau_p^2 u_{ij}) \times \exp(\sigma),$$

which, from Subsection 2.1, leads to $u_{ij} \mid y_{ij}, \mathbf{b}_i \sim \mathrm{GIG}(\frac{1}{2}, \chi_{ij}, \psi)$, where χ_{ij} and ψ are given by

$$\chi_{ij} = \frac{|y_{ij} - \eta_{ij}\left(\boldsymbol{\beta}_p, \mathbf{b}_i\right)|}{\tau_p\sqrt{\sigma}} \quad \text{and} \quad \psi = \frac{\tau_p}{2\sqrt{\sigma}}. \tag{15}$$

From (5), and after generating samples from $f(\mathbf{b}_i \mid \boldsymbol{\theta}^{(k)}, \mathbf{y}_i)$ (see Subsection 3.4), the conditional expectation $\mathrm{E}_{\mathbf{u}_i}[\cdot \mid \boldsymbol{\theta}, \mathbf{b}_i, \mathbf{y}_i]$ in (14) can be computed analytically. Finally, the proposed MCEM algorithm for estimating the parameters of the QR-NLME model can be summarised as follows:

- ***MC E-step:*** Given $\boldsymbol{\theta} = \boldsymbol{\theta}^{(k)}$, for $i = 1, \ldots, n$;
  - **Simulation step:** For $\ell = 1, \ldots, m$, generate $\mathbf{b}_i^{(\ell,k)}$ from $f(\mathbf{b}_i \mid \boldsymbol{\theta}^{(k)}, \mathbf{y}_i)$, as described next in Subsection 3.4.
  - **Monte Carlo approximation:** Using (5) and the $\mathbf{b}_i^{(\ell,k)}$, for $\ell = 1, \ldots, m,$, evaluate

$$\mathrm{E}[g(\mathbf{b}_i, \mathbf{u}_i) \mid \boldsymbol{\theta}^{(k)}, \mathbf{y}_i] \approx \frac{1}{m} \sum_{\ell=1}^{m} \mathrm{E}_{\mathbf{u}_i}[g(\mathbf{b}_i^{(\ell,k)}, \mathbf{u}_i) \mid \boldsymbol{\theta}^{(k)}, \mathbf{b}_i^{(\ell,k)}, \mathbf{y}_i].$$

- ***M-step:*** Update $\widehat{\boldsymbol{\theta}^{(k)}}$ by maximising $Q(\boldsymbol{\theta} \mid \widehat{\boldsymbol{\theta}^{(k)}}) \approx \frac{1}{m}\sum_{l=1}^{m}\sum_{i=1}^{n} \ell_c(\boldsymbol{\theta}; \mathbf{y}_i, \mathbf{b}_i^{(l,k)}, \mathbf{u}_i)$ over $\widehat{\boldsymbol{\theta}^{(k)}}$, leading the following estimates:

$$\begin{aligned}
\widehat{\boldsymbol{\beta}_p}^{(k+1)} &= \widehat{\boldsymbol{\beta}_p}^{(k)} + \left[\sum_{i=1}^{n}\left\{\frac{1}{m}\sum_{\ell=1}^{m}\mathbf{J}_i^{(k)^\top}\mathscr{E}(\boldsymbol{D}_i^{-1})^{(\ell,k)}\mathbf{J}_i^{(k)}\right\}\right]^{-1} \times \\
&\qquad \left[\sum_{i=1}^{n}\left\{\frac{1}{m}\sum_{\ell=1}^{m}\left[2\mathbf{J}_i^{(k)^\top}\mathscr{E}(\boldsymbol{D}_i^{-1})^{(\ell,k)}\left[\mathbf{y}_i - \boldsymbol{\eta}(\widehat{\boldsymbol{\beta}_p}^{(k)}, \mathbf{b}_i^{(\ell,k)}) - \vartheta_p\mathscr{E}(\mathbf{u}_i)^{(\ell,k)}\right]\right]\right\}\right], \\
\widehat{\sigma}^{(k+1)} &= \frac{1}{3N\tau_p^2}\sum_{i=1}^{n}\left\{\frac{1}{m}\sum_{\ell=1}^{m}\left[(\mathbf{y}_i - \boldsymbol{\eta}(\widehat{\boldsymbol{\beta}_p}^{(k+1)}, \mathbf{b}_i^{(\ell,k)}))^\top\mathscr{E}(\boldsymbol{D}^{-1})^{(\ell,k)}(\mathbf{y}_i\boldsymbol{\eta}(\widehat{\boldsymbol{\beta}_p}^{(k+1)}, \mathbf{b}_i^{(\ell,k)}))\right.\right. \\
&\qquad \left.\left. -2\vartheta_p(\mathbf{y}_i\boldsymbol{\eta}(\widehat{\boldsymbol{\beta}_p}^{(k+1)}, \mathbf{b}_i^{(\ell,k)}))^\top\mathbf{1}_{n_i} + \frac{\tau_p^4}{4}\mathscr{E}(\mathbf{u}_i)^{(\ell,k)^\top}\mathbf{1}_{n_i}\right]\right\} \text{ and} \\
\widehat{\boldsymbol{\Psi}}^{(k+1)} &= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{m}\sum_{\ell=1}^{m}\mathbf{b}_i^{(\ell,k)}\mathbf{b}_i^{(\ell,k)^\top}\right],
\end{aligned}$$

  where $\mathbf{J}_i = \partial\boldsymbol{\eta}(\boldsymbol{\beta}_p, \mathbf{b}_i)/\partial\boldsymbol{\beta}_p^\top$, $N = \sum_{i=1}^{n} n_i$ and expressions $\mathscr{E}(\mathbf{u}_i)^{(\ell,k)}$ and $\mathscr{E}(\mathbf{D}_i^{-1})^{(\ell,k)}$ are defined in Appendix B.

Note that, for the MC E-step, we need to generate $\mathbf{b}_i^{(\ell,k)}$, $\ell = 1, \ldots, m$, from $f(\mathbf{b}_i \mid \boldsymbol{\theta}^{(k)}, \mathbf{y}_i)$, where m is the number of Monte Carlo simulations to be used (a number suggested to be large enough). A simulation method to generate samples from $f(\mathbf{b}_i \mid \boldsymbol{\theta}^{(k)}, \mathbf{y}_i)$, is described next in Subsection 3.4.

3.3 A SAEM algorithm

As mentioned in Subsection 2.2, the SAEM circumvents the problem of simulating a large number of latent values at each iteration, leading to a faster and efficient solution to the MCEM algorithm. In summary, the SAEM algorithm proceeds as follows:

- ***E-step:*** Given $\boldsymbol{\theta} = \boldsymbol{\theta}^{(k)}$ for $i = 1, \ldots, n$;
  - **Stochastic approximation:** Update the MC approximations using stochastic approxi-

mations, given by

$$
\begin{aligned}
S_{1,i}^{(k)} &= S_{1,i}^{(k-1)} + \delta_k \left[\frac{1}{m} \sum_{\ell=1}^{m} \mathbf{J}_i^{(k)^\top} \mathscr{E}(\boldsymbol{D}_i^{-1})^{(\ell,k)} \mathbf{J}_i^{(k)} - S_{1,i}^{(k-1)} \right], \\
S_{2,i}^{(k)} &= S_{2,i}^{(k-1)} + \delta_k \left[\frac{1}{m} \sum_{\ell=1}^{m} \left[2\mathbf{J}_i^{(k)^\top} \mathscr{E}(\boldsymbol{D}_i^{-1})^{(\ell,k)} \left[\mathbf{y}_i - \boldsymbol{\eta}(\widehat{\boldsymbol{\beta}_p}^{(k)}, \mathbf{b}_i^{(\ell,k)}) - \vartheta_p \mathscr{E}(\mathbf{u}_i)^{(\ell,k)} \right] \right] - S_{2,i}^{(k-1)} \right], \\
S_{3,i}^{(k)} &= S_{3,i}^{(k-1)} + \delta_k \left\{ \frac{1}{m} \sum_{\ell=1}^{m} \left[(\mathbf{y}_i - \boldsymbol{\eta}(\widehat{\boldsymbol{\beta}_p}^{(k+1)}, \mathbf{b}_i^{(\ell,k)}))^\top \mathscr{E}(\boldsymbol{D}^{-1})^{(\ell,k)} (\mathbf{y}_i - \boldsymbol{\eta}(\widehat{\boldsymbol{\beta}_p}^{(k+1)}, \mathbf{b}_i^{(\ell,k)})) \right. \right. \\
&\qquad \left. \left. -2\vartheta_p (\mathbf{y}_i - \boldsymbol{\eta}(\widehat{\boldsymbol{\beta}_p}^{(k+1)}, \mathbf{b}_i^{(\ell,k)}))^\top \mathbf{1}_{n_i} + \frac{\tau_p^4}{4} \mathscr{E}(\mathbf{u}_i)^{(\ell,k)^\top} \mathbf{1}_{n_i} \right] - S_{3,i}^{(k-1)} \right\} \quad \text{and} \\
S_{4,i}^{(k)} &= S_{4,i}^{(k-1)} + \delta_k \left[\frac{1}{m} \sum_{\ell=1}^{m} [\mathbf{b}_i^{(\ell,k)} \mathbf{b}_i^{(\ell,k)^\top}] - S_{4,i}^{(k-1)} \right].
\end{aligned}
$$

- ***M-step:*** Update $\widehat{\boldsymbol{\theta}^{(k)}}$ by maximizing $Q\left(\boldsymbol{\theta} \mid \widehat{\boldsymbol{\theta}^{(k)}}\right)$ over $\widehat{\boldsymbol{\theta}^{(k)}}$, which leads to the following expressions:

$$
\begin{aligned}
\widehat{\boldsymbol{\beta}_p}^{(k+1)} &= \widehat{\boldsymbol{\beta}_p}^{(k)} + \left[\sum_{i=1}^{n} S_{1,i}^{(k)} \right]^{-1} \sum_{i=1}^{n} S_{2,i}^{(k)}, \\
\widehat{\sigma}^{(k+1)} &= \frac{1}{3N\tau_p^2} \sum_{i=1}^{n} S_{3,i}^{(k)}, \\
\widehat{\Psi}^{(k+1)} &= \frac{1}{n} \sum_{i=1}^{n} S_{4,i}^{(k)}.
\end{aligned} \tag{16}
$$

Given a set of suitable initial values $\widehat{\boldsymbol{\theta}}^{(0)}$ (see Appendix B), the SAEM iterates until convergence at iteration k, if $\max_i \left\{ \frac{|\widehat{\theta}_i^{(k+1)} - \widehat{\theta}_i^{(k)}|}{|\widehat{\theta}_i^{(k)}| + \delta_1} \right\} < \delta_2$, where δ_1 and δ_2 are pre-established small values. As suggested by Searle *et al.* (1992) (page. 269), we use $\delta_1 = 0.001$ and $\delta_2 = 0.0001$. As proposed by Booth & Hobert (1999), we also used a second convergence criteria defined by $\max_i \left\{ \frac{|\widehat{\theta}_i^{(k+1)} - \widehat{\theta}_i^{(k)}|}{\sqrt{\widehat{\mathrm{var}}(\theta_i^{(k)})} + \delta_1} \right\} < \delta_2$.

3.4 Missing data simulation method

In order to generate samples from $f(\mathbf{b}_i \mid \mathbf{y}_i, \boldsymbol{\theta})$, we use the Metropolis-Hastings (MH) algorithm (Metropolis *et al.*, 1953; Hastings, 1970), noting that the conditional distribution $f(\mathbf{b}_i \mid \mathbf{y}_i, \boldsymbol{\theta})$ (omitting $\boldsymbol{\theta}$) can be represented as

$$
f(\mathbf{b}_i \mid \mathbf{y}_i) \propto f(\mathbf{y}_i \mid \mathbf{b}_i) \times f(\mathbf{b}_i),
$$

where $\mathbf{b}_i \sim N_q(\mathbf{0}, \boldsymbol{\Psi})$ and $f(\mathbf{y}_i \mid \mathbf{b}_i) = \prod_{j=1}^{n_i} f(y_{ij} \mid \mathbf{b}_i)$, with $y_{ij} \mid \mathbf{b}_i \sim AL\big(\eta(\mathbf{A}_i\boldsymbol{\beta}_p + \mathbf{B}_i\mathbf{b}_i, \mathbf{x}_{ij}), \sigma, p\big)$. Since the objective function is a product of two distributions (with both support lying in $\mathbb{R}$), a suitable choice for the proposal density is a multivariate normal distribution with mean and variance-covariance matrix given by $\mathrm{E}(\mathbf{b}_i^{(k-1)} \mid \mathbf{y}_i)$ and $\mathrm{Var}(\mathbf{b}_i^{(k-1)} \mid \mathbf{y}_i)$ respectively. These quantities are obtained from the last iteration of the SAEM algorithm. Note that this candidate leads to better acceptance rate, and consequently a faster algorithm.

4 Estimation of the likelihood and standard errors

4.1 Likelihood Estimation

Given the observed data, the likelihood function $\ell_o(\boldsymbol{\theta} \mid \mathbf{y})$ of the model defined in (8)-(9) is given by

$$\ell_o(\boldsymbol{\theta} \mid \mathbf{y}) = \sum_{i=1}^{n} \log f(\mathbf{y}_i \mid \boldsymbol{\theta})) = \sum_{i=1}^{n} \log \int_{\mathbb{R}^q} f(\mathbf{y}_i \mid \mathbf{b}_i; \boldsymbol{\theta})\, f(\mathbf{b}_i; \boldsymbol{\theta})\, d\mathbf{b}_i, \tag{17}$$

where the integral can be expressed as an expectation with respect to $\mathbf{b}_i$, *i.e.*, $E_{\mathbf{b}_i}[f(\mathbf{y}_i \mid \mathbf{b}_i; \boldsymbol{\theta})]$. The evaluation of this integral is not available analytically and is often replaced by its MC approximation involving a large number of simulations. However, alternative importance sampling (IS) procedures might require a smaller number of simulations than the typical MC procedure. Following Meza *et al.* (2012), we can compute this integral using an IS scheme for any continuous distribution $\widehat{f}(\mathbf{b}_i; \boldsymbol{\theta})$ of $\mathbf{b}_i$, having the same support as $f(\mathbf{b}_i; \theta)$. Re-writing (17) as

$$\ell_o(\boldsymbol{\theta} \mid \mathbf{y}) = \sum_{i=1}^{n} \log \int_{\mathbb{R}^q} f(\mathbf{y}_i \mid \mathbf{b}_i; \boldsymbol{\theta})\, \frac{f(\mathbf{b}_i; \boldsymbol{\theta})}{\widehat{f}(\mathbf{b}_i; \boldsymbol{\theta})}\, \widehat{f}(\mathbf{b}_i; \boldsymbol{\theta})\, d\mathbf{b}_i.$$

we can express it as an expectation with respect to $\mathbf{b}_i^*$, where $\mathbf{b}_i^* \sim \widehat{f}(\mathbf{b}_i^*; \boldsymbol{\theta})$. Thus, the likelihood function can now be expressed as

$$\ell_o(\boldsymbol{\theta} \mid \mathbf{y}) \approx \sum_{i=1}^{n} \log \left\{ \frac{1}{m} \sum_{\ell=1}^{m} \left[\prod_{j=1}^{n_i} [f(y_{ij} \mid \mathbf{b}_i^{*(\ell)}; \boldsymbol{\theta})]\, \frac{f(\mathbf{b}_i^{*(\ell)}; \boldsymbol{\theta})}{\widehat{f}(\mathbf{b}_i^{*(\ell)}; \boldsymbol{\theta})} \right] \right\}, \tag{18}$$

where $\{\mathbf{b}_i^{*(\ell)}\}$, $l = 1, \ldots, m$, is an MC sample from $\widehat{f}(\mathbf{b}_i^*; \boldsymbol{\theta})$, and $f(\mathbf{y}_i \mid \mathbf{b}_i^{*(\ell)}; \boldsymbol{\theta})$ is expressed as $\prod_{j=1}^{n_i} f(y_{ij} \mid \mathbf{b}_i^{*(\ell)}; \boldsymbol{\theta})$ due to the conditional independence assumption. An efficient choice for $\widehat{f}(\mathbf{b}_i^{*(\ell)}; \theta)$ is $f(\mathbf{b}_i \mid \mathbf{y}_i)$. Therefore, we use the same proposal distribution discussed in Subsection 3.4, generating $\mathbf{b}_i^{*(\ell)} \sim N_q(\widehat{\boldsymbol{\mu}}_{\mathbf{b}_i}, \widehat{\boldsymbol{\Sigma}}_{\mathbf{b}_i})$, where $\widehat{\boldsymbol{\mu}}_{\mathbf{b}_i} = \mathrm{E}(\mathbf{b}_i^{(w)} \mid \mathbf{y}_i)$ and $\widehat{\boldsymbol{\Sigma}}_{\mathbf{b}_i} = \mathrm{Var}(\mathbf{b}_i \mid \mathbf{y}_i)$, which are estimated empirically during the last few iterations of the SAEM algorithm at convergence.

4.2 Standard error approximation

Louis' missing information principle (Louis, 1982) relates the score function of the incomplete data log-likelihood with the complete data log-likelihood through the conditional expectation $\boldsymbol{\nabla}_o(\boldsymbol{\theta}) = \mathrm{E}_{\boldsymbol{\theta}}[\boldsymbol{\nabla}_c(\boldsymbol{\theta}; \mathbf{Y}_{com} | \mathbf{Y}_{obs})]$, where $\boldsymbol{\nabla}_o(\theta) = \partial \ell_o(\boldsymbol{\theta}; \mathbf{Y}_{obs})/\partial\theta$ and $\boldsymbol{\nabla}_c(\boldsymbol{\theta}) = \partial \ell_c(\theta; \mathbf{Y}_{com})/\partial\boldsymbol{\theta}$ are the

score functions for the incomplete and complete data, respectively. As defined in Meilijson (1989), the empirical information matrix can be computed as

$$\mathbf{I}_e(\boldsymbol{\theta}|\mathbf{y}) = \sum_{i=1}^{n} \mathbf{s}(\mathbf{y}_i|\boldsymbol{\theta})\,\mathbf{s}^{\top}(\mathbf{y}_i|\widehat{\boldsymbol{\theta}}) - \frac{1}{n}\mathbf{S}(\mathbf{y}|\boldsymbol{\theta})\,\mathbf{S}^{\top}(\mathbf{y}|\boldsymbol{\theta}), \tag{19}$$

where $\mathbf{S}(\mathbf{y}|\boldsymbol{\theta}) = \sum_{i=1}^{n} \mathbf{s}(\mathbf{y}_i|\boldsymbol{\theta})$, with $\mathbf{s}(\mathbf{y}_i|\boldsymbol{\theta})$ the empirical score function for the i-th individual. Replacing $\boldsymbol{\theta}$ by its ML estimator $\hat{\boldsymbol{\theta}}$ and considering $\nabla_o(\hat{\boldsymbol{\theta}}) = \mathbf{0}$, equation (19) takes the simple form

$$\mathbf{I}_e(\widehat{\boldsymbol{\theta}}|\mathbf{y}) = \sum_{i=1}^{n} \mathbf{s}(\mathbf{y}_i|\widehat{\boldsymbol{\theta}})\,\mathbf{s}^{\top}(\mathbf{y}_i|\widehat{\boldsymbol{\theta}}). \tag{20}$$

At the kth iteration, the empirical score function for the i-th subject can be computed as

$$\mathrm{s}(\mathbf{y}_i|\boldsymbol{\theta})^{(k)} = \mathbf{s}(\mathbf{y}_i|\boldsymbol{\theta})^{(k-1)} + \delta_k \left[\frac{1}{m} \sum_{\ell=1}^{m} \mathbf{s}(\mathbf{y}_i, \mathbf{q}^{(\ell,k)}; \boldsymbol{\theta}^{(k)}) - \mathbf{s}(\mathbf{y}_i|\boldsymbol{\theta})^{(k-1)} \right], \tag{21}$$

where $\mathbf{q}^{(\ell,k)}$, $\ell = 1,\ldots,m$, are the simulated missing values drawn from the conditional distribution $f(\cdot|\theta^{(k-1)}, \mathbf{y}_i)$. Thus, at iteration k, the observed information matrix can be approximated as $\mathbf{I}_e(\boldsymbol{\theta}|\mathbf{y})^{(k)} = \sum_{i=1}^{n} \mathbf{s}(\mathbf{y}_i|\boldsymbol{\theta})^{(k)}\,\mathbf{s}^{\top}(\mathbf{y}_i|\boldsymbol{\theta})^{(k)}$, such that at convergence, $\mathbf{I}_e^{-1}(\widehat{\boldsymbol{\theta}}|\mathbf{y}) = (\mathbf{I}_e(\boldsymbol{\theta}|\mathbf{y})|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}})^{-1}$ is an estimate of the covariance matrix of the parameter estimates. Expressions for the elements of the score vector with respect to $\boldsymbol{\theta}$ are given in Appendix C.

5 Simulated data

In order to examine the performance of the proposed method, we conduct some simulation studies. The first simulation study shows that the ML estimates based on the SAEM algorithm provide good asymptotic properties. The second study investigates the consequences in population inferences when the normality assumption is inappropriate. In order to do that, we used a heavy tailed distribution for the random error term, testing the robustness of the proposed method in terms of the parameter estimation.

5.1 Asymptotic properties

As in Pinheiro & Bates (1995), we performed the first simulation study with the following three parameter non-linear growth-curve logistic model:

$$y_{ij} = \frac{\beta_1 + b_{1i}}{1 + \exp\left(-[t_{ij} - \beta_2]/\beta_3\right)} + \varepsilon_{ij},\ i = 1,\ldots,n, \quad ,j = 1,\ldots,10, \tag{22}$$

where $t_{ij} = 100, 267, 433, 600, 767, 933, 1100, 1267, 1433, 1600$ for all i. The goal is to estimate the fixed effects parameters β's for a grid of percentiles $p = \{0.50, 0.75, 0.95\}$. A random effect b_{1i}, for $i = 1,\ldots,n$ is added to the first growth parameter β_1 and its effect over the growth-curve is shown in Figure 4.

Parameters interpretation for this model is discussed in Section 6. The random effect b_{1i} and the error term $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}\ldots,\varepsilon_{i10})^{\top}$ are non-correlated. In fact, $b_{1i} \overset{\text{iid}}{\sim} \mathrm{N}(0,\sigma_b^2)$ and $\varepsilon_{ij} \overset{\text{iid}}{\sim} \mathrm{AL}(0,\sigma_e,p)$.

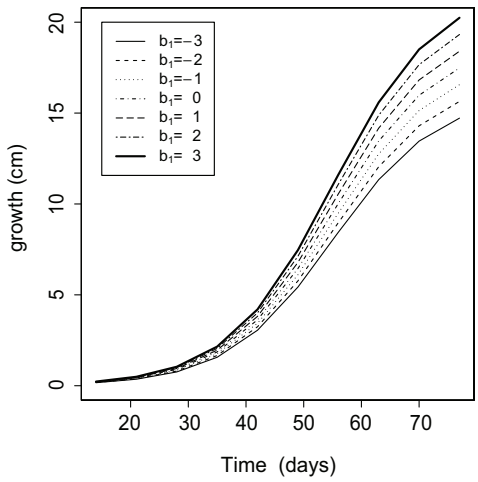

Figure 2: Effect of including a random effect b_1 in the first parameter of the non-linear growth-curve logistic model.

We set $\boldsymbol{\beta}_p = (\beta_1, \beta_2, \beta_3)^\top = (200, 700, 350)^\top$, $\sigma_e = 0.5$ and $\sigma_b^2 = 10$. Using the notation in (7), the matrices $\mathbf{A}_i$ and $\mathbf{B}_i$ are given by $\mathbf{I}_3$ and $(1,0,0)^\top$ respectively. For different sample sizes $n =$ 25, 50, 100 and 200, we generate 100 data samples for each scenario. In addition, we choose $m =$ 20, $c = 0.25$ and $W = 500$ for the SAEM algorithm convergence parameters. Note, the choice of c depends on the dataset, and also the underlying model. We set $c = 0.25$, given that an initial run of 125 iterations (which is 25% of W) for the 0.05th quantile led to convergence to the neighborhood solution. For all scenarios, we compute the square root of the mean square error (RMSE), bias (Bias) and Monte Carlo standard deviation (MC-Sd) for each parameter over the 100 replicates. These quantities are defined as

$$\text{MC-Sd}(\widehat{\theta}_i) = \sqrt{\frac{1}{99}\sum_{j=1}^{100}\left(\widehat{\theta}_i^{(j)} - \overline{\widehat{\theta}_i}\right)^2} \qquad \text{and} \qquad \text{Bias}(\widehat{\theta}_i) = \overline{\widehat{\theta}_i} - \theta_i \tag{23}$$

where $\text{RMSE}(\widehat{\theta}_i) = \sqrt{\text{MC-Sd}^2(\widehat{\theta}_i) + \text{Bias}^2(\widehat{\theta}_i)}$, the Monte Carlo mean $\overline{\widehat{\theta}_i} = \frac{1}{100}\sum_{j=1}^{100}\widehat{\theta}_i^{(j)}$ (MC Mean) and $\theta_i{}^{(j)}$ is the estimate of θ_i from the j-th sample, $j = 1 \ldots 100$. Based on Figure 3, we conclude that the bias in the estimation of fixed effects converge to zero when n increases.

Table 1: Simulation 1: Monte Carlo mean and standard deviation (MC Mean and MC-Sd) for the fixed effects $\boldsymbol{\beta}$ and scale parameter σ_e obtained after fitting the QR-NLME model under different settings of quantiles and sample sizes. Results based on 100 simulated samples.

		β_1		β_2		β_3		σ_e	
Quantile (%)	n	MC Mean	MC-Sd	MC Mean	MC-Sd	MC Mean	MC-Sd	MC Mean	MC-Sd
50	25	199.75	(2.35)	700.19	(2.00)	350.13	(1.35)	0.503	(0.035)
	50	199.79	(1.69)	700.09	(1.29)	350.03	(0.86)	0.498	(0.021)
	100	200.16	(1.15)	700.08	(0.92)	350.06	(0.72)	0.497	(0.017)
	200	200.03	(0.75)	699.96	(0.64)	349.98	(0.50)	0.499	(0.012)
75	25	203.77	(2.50)	700.18	(2.07)	350.15	(1.56)	0.499	(0.035)
	50	203.90	(1.81)	700.20	(1.60)	350.16	(1.11)	0.495	(0.025)
	100	204.20	(1.31)	699.83	(1.08)	349.88	(0.74)	0.499	(0.017)
	200	204.34	(0.92)	700.00	(0.70)	350.01	(0.49)	0.498	(0.011)
95	25	201.15	(2.79)	700.26	(6.52)	350.14	(3.92)	0.506	(0.035)
	50	201.77	(2.15)	700.53	(4.84)	349.74	(2.83)	0.508	(0.024)
	100	201.94	(1.56)	700.18	(3.55)	349.73	(2.32)	0.505	(0.015)
	200	202.11	(1.08)	700.06	(2.60)	349.98	(1.54)	0.502	(0.012)

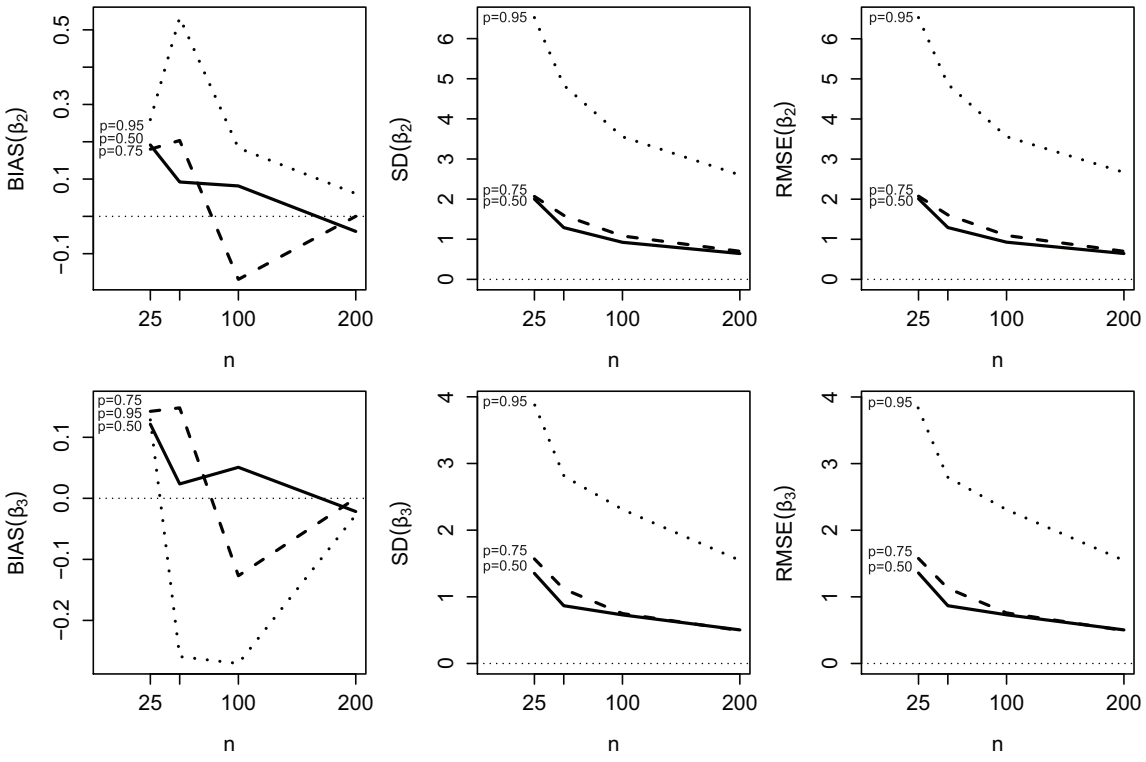

Figure 3: Bias, Standard Deviation and RMSE for β_1 (upper panel) and β_2 (lower panel) for different sample sizes over the quantiles $p = \{0.50, 0.90, 0.95\}$.

The values of MC-Sd and RMSE decrease monotonically when n is increased. Note that for quantile $q = 95$, the standard deviation is much higher than quantiles $q = 50$ and $q = 75$. As a

conclusion, we can say that, in general, the bias and MSE converge to zero when the sample size is increasing, indicating that proposed SAEM algorithm provides good asymptotic properties for the ML estimation.Table 1 also shows the estimation of σ_e. In this case, small standard deviations and good asymptotic properties in terms of bias and SD are observed.

5.2 Robustness study

The aim of this simulation is to study the behaviour of parameter estimates when the distribution of random effects is misspecified. We consider a similar simulation scheme as in the previous subsection, but considering a set of quantiles $\{0.50, 0.75\}$ and a fixed sample size $n = 50$. We consider 100 Monte Carlo samples, generating the random effect term from (a) a Student's-t distribution with $\nu = 4$ degrees of freedom and from (b) a contaminated normal distribution with parameters $\nu_1 = 0.1$ and $\nu_2 = \{0.1, 0.2, 0, 3\}$), *i.e.*, three scenarios of contamination, say, 10%, 20% and 30%. We set the value of parameters as follows: $\boldsymbol{\beta}_p = (200, 700, 350)^\top$, $\sigma_e = 0.5$ and $\sigma_b^2 = 10$.

From Table 2 we can see that the proposed model is robust even when the level of contamination is high. For quantile 0.75, the parameter β_1 tends to increase for higher levels of contamination. As expected, the MC-Sd and RMSE increase when the distribution of the random effects is heavy-tailed.

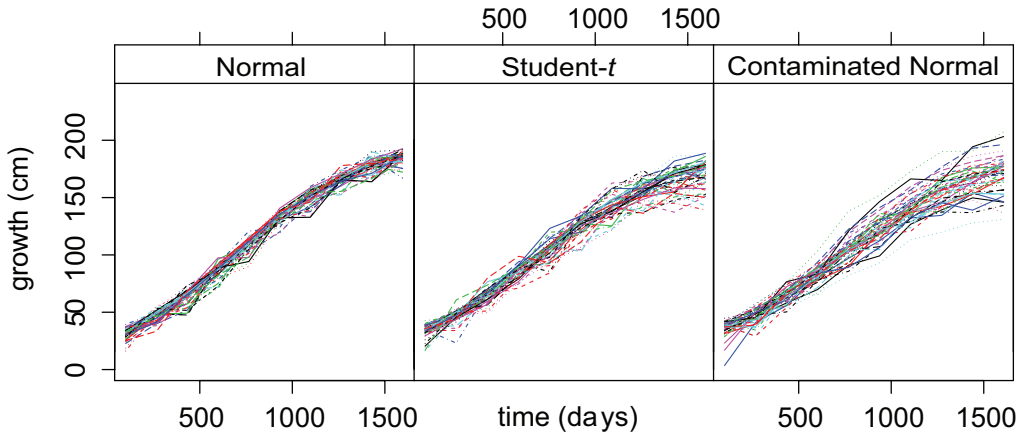

Figure 4: 50 simulated curves from the growth-curve logistic model using different distributions for the random effect term: normal (right), Student's-t with $\nu = 4$ (center), contaminated normal with $\nu_1 = 0.1$ and $\nu_2 = 0.1$ (left). In all cases the location and scale parameters are $\mu = 0$ and $\sigma_b^2 = 10$ respectively.

6 Illustrative examples

In this section, we illustrate the application of our method analysing two longitudinal datasets.

6.1 Growth curve: Soybean data

For the first application, we consider the Soybean genotypes data analysed previously by Davidian & Giltinan (1995) and Pinheiro & Bates (2000). The experiment consists in measuring (along the time) the leaf weight (in g) as a measure of growth of two kinds of Soybean genotype plants, namely, a commercial variety called Forrest (F), and an experimental strain called Plan Introduction #416937 (P). The samples were taken approximately weekly during 8 to 10 weeks. For three consecutive years, say, 1988, 1989 and 1990, the plants were planted in 16 plots (8 per each genotype) and the mean leaf weight of six randomly selected plants was measured.

We use the three parameter logistic model in (22) introducing a random effect term for each parameter and a dichotomic covariate (*gen*) as

$$y_{ij} = \frac{\varphi_{1i}}{1+\exp\left(-[t_{ij}-\varphi_{2i}]/\varphi_{3i}\right)} + \varepsilon_{ij},\ i = 1,\ldots,412,\ \ j = 1,\ldots,n_i, \tag{24}$$

where, $\varphi_{1i} = \beta_1 + \beta_4 gen_i + b_{1i}$, $\varphi_{2i} = \beta_2 + b_{2i}$ and $\varphi_{3i} = \beta_3 + b_{3i}$. The observed value y_{ij} represents the mean weight of leaves (in g) from six randomly selected soybean plants in the ith plot, after t_{ij} days of planted; gen_i is a dichotomic variable indicating the genotype of the plant i (0=forrest, 1=plan Introduction) and ε_{ij} is the measurement error term. Moreover, $\boldsymbol{\beta}_p = (\beta_1,\beta_2,\beta_3,\beta_4)^\top$ and $\mathbf{b}_i = (b_{1i},b_{2i},b_{3i})^\top$ are the fixed and random effects vector respectively. The matrices $\mathbf{A}_i$ and $\mathbf{B}_i$ are

Table 2: Simulation 2: MC Mean, Bias, MC-Sd and RMSE for the fixed effects $\boldsymbol{\beta}$ and scale parameter σ_e obtained after fitting the QR-NLME model for quantiles 0.50 and 0.75 using four different distribution settings for the random effects. Results based on 100 simulated samples.

Fit		Quantile 50%				Quantile 75%			
		β_1 (200)	β_2 (700)	β_3 (350)	σ_e (0.5)	β_1 (200)	β_2 (700)	β_3 (350)	σ_e (0.5)
Student-t_4	MC Mean	200.22	700.00	349.99	0.501	204.43	700.39	350.18	0.501
	Bias	0.22	0.00	-0.01	0.001	4.43	0.39	0.18	0.001
	MC-Sd	(1.98)	(1.28)	(0.98)	(0.024)	(2.17)	(1.69)	(1.09)	(0.024)
	RMSE	1.99	1.28	0.98	0.024	4.93	1.74	1.11	0.024
Contamination									
10%	MC Mean	199.87	700.10	349.9	0.499	205.02	700.18	350.05	0.501
	Bias	-0.13	0.10	-0.1	-0.001	5.02	0.18	0.05	0.001
	MC-Sd	(1.90)	(1.26)	(0.88)	(0.024)	(1.92)	(1.80)	(1.16)	(0.024)
	RMSE	1.90	1.27	0.88	0.024	5.38	1.81	1.16	0.024
20%	MC Mean	200.05	699.91	350.08	0.497	205.35	700.20	350.11	0.496
	Bias	0.05	-0.09	0.08	-0.003	5.35	0.20	0.11	-0.004
	MC-Sd	(1.96)	(1.28)	(0.90)	(0.024)	(2.00)	(1.55)	(1.19)	(0.023)
	RMSE	1.96	1.28	0.90	0.024	5.71	1.56	1.20	0.023
30%	MC Mean	200.16	700.06	350.07	0.496	206.63	699.91	350.01	0.497
	Bias	0.16	0.06	0.07	-0.004	6.63	-0.09	0.01	-0.003
	MC-Sd	(2.10)	(1.05)	(0.93)	(0.024)	(2.60)	(1.60)	(1.06)	(0.022)
	RMSE	2.11	1.05	0.93	0.024	7.13	1.60	1.06	0.023

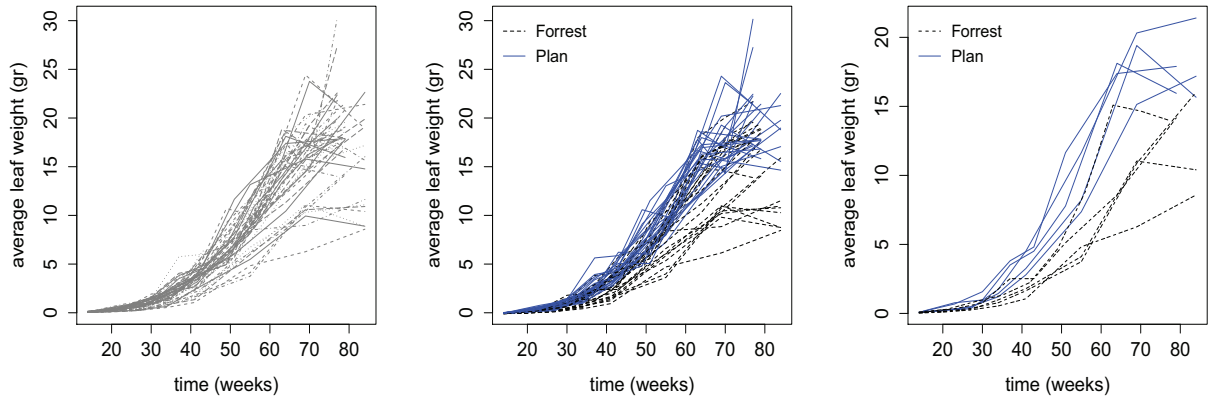

Figure 5: Soybean data: (a) Leaf weight profiles versus time. (b) Leaf weight profiles versus time by genotype. (c) Ten randomly selected leaf weight profiles versus time been five per each genotype.

defined as

$$\mathbf{A}_i = \begin{pmatrix} 1 & 0 & 0 & gen_i \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \qquad \text{and} \qquad \mathbf{B}_i = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{25}$$

The three parameter interpretation are the asymptotic leaf weight, the time at which the leaf reaches half of its asymptotic weight and the time elapsed between the leaf reaching half and $0.7311 = 1/(1+e^{-1})$ of its asymptotic weight, respectively. Since the aim of the study is to compare the final (asymptotic) growth of the two kind of Soybeans, the covariate gen_i was incorporated in the first component of the growth function. Therefore, the coefficient β_4 represents the difference (in g) of the asymptotic leaf weight between the plan introduction type and forrest one (control). Figure 5 shows the leaf weight profiles.

Figure 6 shows the fitted regression lines for quantiles 0.10, 0.25, 0.50, 0.75 and 0.90 by genotype. From this figure we can see how the extreme quantiles estimation functions captures the full data variability, detecting some atypical observations (particularly for the Plan Introduction genotype).

Figure 9 in Appendix D shows a summary of the obtained results. We can see that the effect of the genotype results significant for all the quantile profile. Moreover, the difference varies with respect to the conditional quantile been more significant for lower quantiles. Using the information provided by the 95th percentile, we conclude that the Soybean plants growing more have a mean leaf weight around 19.35 grams for the Forrest genotype and 23.25 grams for the Plan Introduction genotype, then the asymptotic difference for the two genotypes is around 4 grams. Finally, it is important to stress that the convergence of the fixed effect estimates and variance components is analysed using a graphical criteria as is shown in Figure 11 (Appendix D).

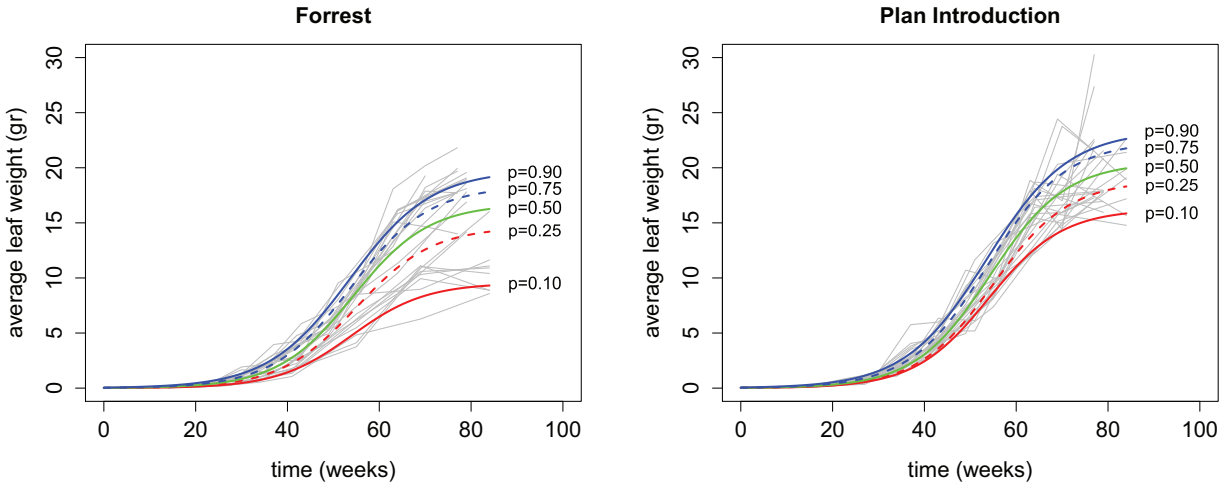

Figure 6: Soybean data: Fitted quantile regression for several quantiles.

6.2 HIV viral load study

The data set belongs to a clinical trial (ACTG 315) studied in previous works by Wu (2002) and Lachos *et al.* (2013). In this study, the HIV viral load of 46 HIV-1 infected patients under antiretroviral treatment (protease inhibitor and reverse transcriptase inhibitor drugs) is analysed. The viral load and some other covariates were measured several times after the start of treatment. Wu (2002) found that the only significance covariate for modelling the virus load was the CD4. Figure 7 shows the profile of viral load in log10 scale and CD4 cell count/100 per cubic ml versus time (in days/100) for six randomly selected patients. We can see that there exist some inverse relationship between the viral load and the CD4 cell count, *i.e.*, high CD4 cell count leads to lower levels of viral load. This is because the CD4 cells (also called T-cells) alert the immune system in the case of an invasion of viruses and/or bacteria. Consequently, lower CD4 count means a weaker immune system.

In order to fit the ACTG 315 data, we propose a bi-phasic non-linear model considered by Wu (2002) and also used by Lachos *et al.* (2013). The proposed NLME model is given by:

$$y_{ij} = \log_{10}\left(e^{(\varphi_{1i}-\varphi_{2i}t_{ij})} + e^{(\varphi_{3i}-\varphi_{4i}t_{ij})}\right) + \varepsilon_{ij}, \quad i = 1,\ldots,46, \;\; j = 1,\ldots,n_i, \tag{26}$$

with $\varphi_{1i} = \beta_1 + b_{1i}$, $\varphi_{2i} = \beta_2 + b_{2i}$, $\varphi_{3i} = \beta_3 + b_{3i}$, $\varphi_{4ij} = \beta_4 + \beta_5 CD4_{ij} + b_{4i}$, where the observed value y_{ij} represents the log-10 transformation of the viral load for the ith patient at time j, $CD4_{ij}$ is the CD4 cell count (in cells/100mm^3) for the ith patient at time j and ε_{ij} is the measurement error term. As in the previous case, $\boldsymbol{\beta}_p = (\beta_1,\beta_2,\beta_3,\beta_4,\beta_5)^\top$ and $\mathbf{b}_i = (b_{1i},b_{2i},b_{3i},b_{4i})^\top$ denotes the fixed and random effects vector respectively, and $\mathbf{CD4}_i = (CD4_{i1},\ldots,CD4_{in_i})^\top$. The matrices $\mathbf{A}_i$ and $\mathbf{B}_i$ are defined as

$$\mathbf{A}_i = \begin{pmatrix} \mathbf{I}_3 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_{n_i} & \mathbf{CD4}_i \end{pmatrix} \qquad \text{and} \qquad \mathbf{B}_i = \begin{pmatrix} \mathbf{I}_3 & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_{n_i} \end{pmatrix}. \tag{27}$$

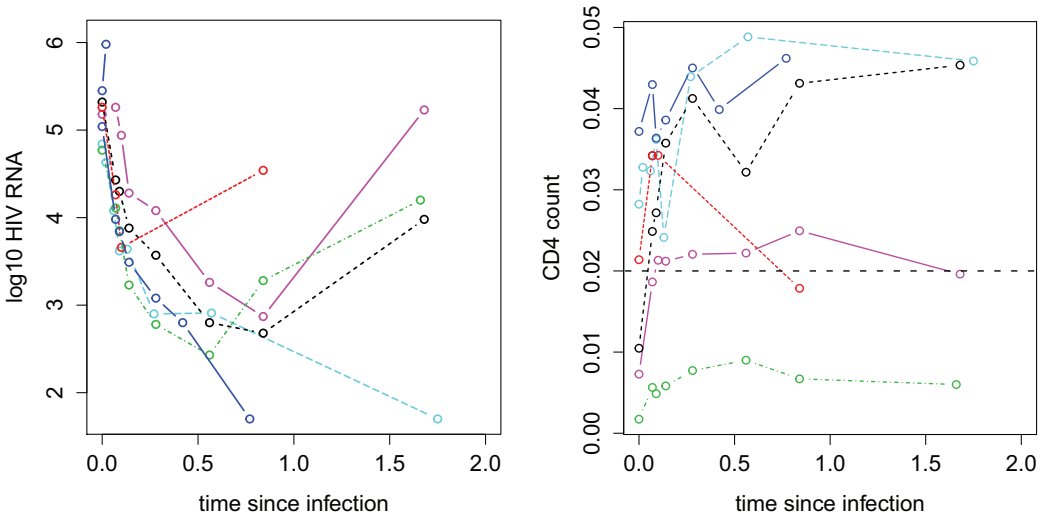

Figure 7: ACTG 315 data. Profiles of viral load (response) in log10 scale and CD4 cell count (in cells/100mm^3) for six randomly selected patients.

The parameters φ_{2i} and φ_{4i} are the two-phase viral decay rates, which represent the minimum turnover rates of productively infected cells and that of latently or long-lived infected cells if therapy was successful, respectively. For more details about the model in (26) see Grossman *et al.* (1999) and Perelson *et al.* (1997).

Figure 8 shows the fitted regression lines for quantiles 0.10, 0.25, 0.50, 0.75 and 0.90 for the ACTG 315 data. In order to plot this, first, we have fixed the CD4 covariate using the predicted sequence from a linear regression (including a quadratic term) for explaining the CD4 cell count along the time. We can see how quantile estimated functions follow the data behaviour and turn easily to estimate a specific viral load quantile at any time of the experiment. Extreme quantile functions bound the most of the observed profiles and evidence possible influential observations.

The results after fitting QR-NLME model over the grid of quantiles $p = \{0.05, 0.10, \ldots, 0.95\}$ are shown in figure 10 in Appendix D. The first phase viral decay rate is positive and its effect tends to increase proportionally along quantiles. Moreover, the second phase viral decay rate is positive correlated with the CD4 count and therefore with the duration of therapy. Consequently, more days of treatment implies a higher CD4 cell count and therefore a higher second phase viral decay. The CD4 cell process for this model has a different behaviour than for the expansion phase (Huang & Dagne (2011)). The significance of the CD4 covariate increases positively with respect to quantiles (until quantile $p = 0.60$ approximately) and its effect becomes constant for greater quantiles. As in the previous case, the convergence of estimates for all parameters were also assessed using the graphical criteria in Figure 12 in Appendix D.

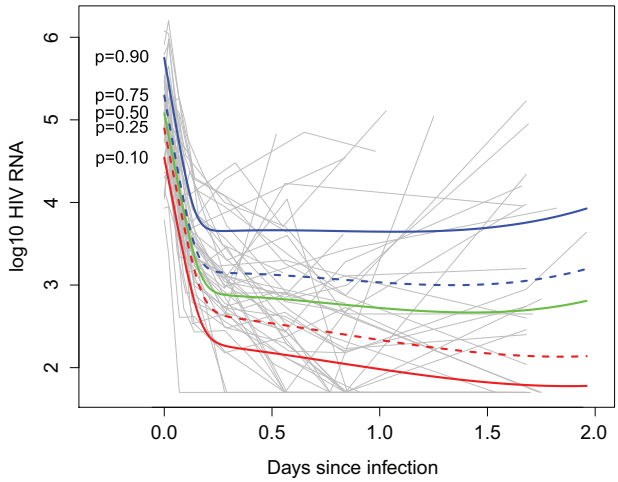

Figure 8: ACTG 315 data: Fitted quantile regression functions.

7 Conclusions

In this paper, we investigate quantile regression under non-linear mixed effects models from a likelihood-based perspective. The AL distribution and SAEM algorithm are combined efficiently to propose an exact ML estimation method, in contrast to the approximated method proposed by Geraci & Bottai (2014) for LMM. We evaluate the robustness of estimates, as well as the finite sample performance of the algorithm and the asymptotic properties of the ML estimates through empirical experiments. To the best of our knowledge, we consider that this paper is the first attempt for exact ML estimation in the context of QR-NLME models. The methods developed can be readily implemented inside `R` through package `qrNLMM()`, making our approach quite powerful and accessible to practitioners.

Certainly, other distributions can be used as alternatives to the AL distribution. Recently, Wichitaksorn *et al.* (2014) presented a generalized class of skew density for QR that provides competing solutions to the AL distribution-based formulation. However, their exploration is limited to the simple linear QR framework. Also, due to the lack of a relevant stochastic representation, the corresponding EM-type implementation can lead to difficulties. Recently, Galarza, C.E. and Benites, L.E. and Lachos, Victor V.H. (2015) presented an `R` package for a linear QR using a new family of skew distributions that includes the ones formulated in Wichitaksorn *et al.* (2014) as special cases. This family includes the skewed version of Normal, Student-*t*, Laplace, Contaminated Normal and Slash distribution, all with the zero quantile property for the error term, and with a convenient stochastic representation. Undoubtedly, incorporating this skewed class into our NLMM proposition can enhance flexibility, and potentially improve our inference.

Also, for modelling both skewness and heavy tails in the random effects, the use of scale mixtures of skew-normal (SMSN) distributions (Lachos *et al.*, 2010) is a feasible choice. Also, HIV viral loads studies include covariates (CD4 cell counts) that often comes with substantial measurement errors (Wu, 2002). How to incorporate measurement error in covariates within our

robust framework can also be part of future research. An in-depth investigation of such extensions is beyond the scope of the present paper, but certainly an interesting topic for future research.

Acknowledgements

The research of V. H. Lachos was supported by Grant 306334/2015-1 from Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq-Brazil) and by Grant 2014/02938-9 from Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP-Brazil).

Appendix A Specification of initial values

It is well known that a smart choice of the initial values for the ML estimates can assure a fast convergence of an algorithm to the global maxima solution. Obviating the random effects term, *i.e.*, $\mathbf{b}_i = \mathbf{0}$, let $\mathbf{y}_i \sim \mathrm{AL}(\boldsymbol{\eta}(\boldsymbol{\beta}_p, \mathbf{0}), \sigma, p)$. Next, considering the ML estimates for $\boldsymbol{\beta}_p$ and σ as defined in Yu & Zhang (2005) for this model, we follow the steps below for the QR-LME model implementation:

1. Compute an initial value $\widehat{\boldsymbol{\beta}}_p^{(0)}$ as

$$\widehat{\boldsymbol{\beta}}_p^{(0)} = \arg\min_{\beta_p \in \mathbb{R}^k} \sum_{i=1}^{n} \rho_p(\mathbf{y}_i - \boldsymbol{\eta}(\boldsymbol{\beta}_p, \mathbf{0})).$$

2. Using the initial value for $\widehat{\boldsymbol{\beta}}_p^{(0)}$ obtained above, compute $\widehat{\sigma}^{(0)}$ as

$$\widehat{\sigma}^{(0)} = \frac{1}{n} \sum_{i=1}^{n} \rho_p(\mathbf{y}_i - \boldsymbol{\eta}(\boldsymbol{\beta}_p, \mathbf{0})).$$

3. Use a $q \times q$ identity matrix $\mathbf{I}_{q\times q}$ for the the initial value $\boldsymbol{\Psi}^{(0)}$.

Appendix B Computing the conditional expectations

Due the independence between $u_{ij}\big|\, y_{ij}, \mathbf{b}_i$ and $u_{ik}\big|\, y_{ik}, \mathbf{b}_i$, for all $j,k = 1,2,\ldots,n_i$ and $j \neq k$, we can write $\mathbf{u}_i \mid \mathbf{y}_i, \mathbf{b}_i = [\ u_{i1}|y_{i1}, \mathbf{b}_i \quad u_{i2}|y_{i2}, \mathbf{b}_i \quad \cdots \quad u_{in_i}|y_{in_i}, \mathbf{b}_i\ ]^\top$. Using this fact, we are able to compute the conditional expectations $\mathscr{E}(\mathbf{u}_i)$ and $\mathscr{E}(\mathbf{D}_i^{-1})$ in the following way. Using matrix expectation properties, we define these expectations as

$$\mathscr{E}(\mathbf{u}_i) = [\mathscr{E}(u_{i1})\ \mathscr{E}(u_{i1})\ \cdots\ \mathscr{E}(u_{in_i})]^\top \tag{B.1}$$

and

$$\mathscr{E}(\mathbf{D}_i^{-1}) = \mathrm{diag}(\mathscr{E}(\mathbf{u}_i^{-1})) = \begin{bmatrix} \mathscr{E}(u_{i1}^{-1}) & 0 & \ldots & 0 \\ 0 & \mathscr{E}(u_{i2}^{-1}) & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \mathscr{E}(u_{in_i}^{-1}) \end{bmatrix}. \tag{B.2}$$

We already have $u_{ij} \mid y_{ij}, \mathbf{b}_i \sim \text{GIG}(\frac{1}{2}, \chi_{ij}, \psi)$ where χ_{ij} and ψ are defined in (15). Then, using (5), we compute the moments involved in the equations above as $\mathscr{E}(u_{ij}) = \frac{\chi_{ij}}{\psi}(1 + \frac{1}{\chi_{ij}\psi})$ and $\mathscr{E}(u_{ij}^{-1}) = \frac{\psi}{\chi_{ij}}$. Thus, for iteration k of the algorithm and for the ℓth Monte Carlo realization, we can compute $\mathscr{E}(\mathbf{u}_i)^{(\ell,k)}$ and $\mathscr{E}[\mathbf{D}_i^{-1}]^{(\ell,k)}$ using equations (B.1)-(B.2) where

$$\mathscr{E}(u_{ij})^{(\ell,k)} = \frac{2|y_{ij} - \eta_{ij}(\boldsymbol{\beta}_p^{(k)}, \mathbf{b}_i^{(\ell,k)})| + 4\sigma^{(k)}}{\tau_p^2} \qquad \text{and} \quad \mathscr{E}(u_{ij}^{-1})^{(\ell,k)} = \frac{\tau_p^2}{2|y_{ij} - \eta_{ij}(\boldsymbol{\beta}_p^{(k)}, \mathbf{b}_i^{(\ell,k)})|}.$$

Appendix C The empirical information matrix

In light of (11), the complete log-likelihood function can be rewritten as

$$\ell_{ci}(\boldsymbol{\theta}) = -\frac{3}{2}n_i \log\sigma - \frac{1}{2\sigma\tau_p^2}\zeta_i^\top \mathbf{D}_i^{-1}\zeta_i - \frac{1}{2}\log|\boldsymbol{\Psi}| - \frac{1}{2}\mathbf{b}_i^\top \boldsymbol{\Psi}^{-1}\mathbf{b}_i - \frac{1}{\sigma}\mathbf{u}_i^\top \mathbf{1}_{n_i} \tag{C.1}$$

where $\zeta_i = \mathbf{y}_i - \boldsymbol{\eta}(\boldsymbol{\beta}_p, \mathbf{b}_i) - \vartheta_p \mathbf{u}_i$ and $\boldsymbol{\theta} = (\boldsymbol{\beta}_p^\top, \sigma, \boldsymbol{\alpha}^\top)^\top$. Differentiating with respect to $\boldsymbol{\theta}$, we have the following score functions:

$$\frac{\partial \ell_{ci}(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}_p} = \frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\beta}_p}\frac{\partial \zeta_i}{\partial \boldsymbol{\eta}}\frac{\partial \ell_{ci}(\boldsymbol{\theta})}{\partial \zeta_i} = \frac{1}{\sigma\tau_p^2}\mathbf{J}_\mathbf{i}^\top \mathbf{D}_i^{-1}\zeta_i,$$

with $\mathbf{J_i}$ defined in section 3.2. and

$$\frac{\partial \ell_{ci}(\boldsymbol{\theta})}{\partial \sigma} = -\frac{3n_i}{2}\frac{1}{\sigma} + \frac{1}{2\sigma^2\tau_p^2}\zeta_i^\top \mathbf{D}_i^{-1}\zeta_i + \frac{1}{\sigma^2}\mathbf{u}_i^\top \mathbf{1}_{n_i}.$$

Let $\boldsymbol{\alpha}$ be the vector of reduced parameters from $\boldsymbol{\Psi}$, the dispersion matrix for $\mathbf{b}_i$. Using the trace properties and differentiating the complete log-likelihood function, we have that

$$\begin{aligned}
\frac{\partial \ell_{ci}(\boldsymbol{\theta})}{\partial \boldsymbol{\Psi}} &= \frac{\partial}{\partial \boldsymbol{\Psi}}\left[-\frac{n}{2}log|\boldsymbol{\Psi}| - \frac{1}{2}\text{tr}\{\boldsymbol{\Psi}^{-1}\mathbf{b}_i\mathbf{b}_i^\top\}\right] \\
&= -\frac{1}{2}\text{tr}\{\boldsymbol{\Psi}^{-1}\} + \frac{1}{2}\text{tr}\{\boldsymbol{\Psi}^{-1}\boldsymbol{\Psi}^{-1}\mathbf{b}_i\mathbf{b}_i^\top\} \\
&= \frac{1}{2}\text{tr}\{\boldsymbol{\Psi}^{-1}(\mathbf{b}_i\mathbf{b}_i^\top - \boldsymbol{\Psi})\boldsymbol{\Psi}^{-1}\}
\end{aligned}$$

Next, taking derivatives with respect to a specific α_j from $\boldsymbol{\alpha}$ based on the chain rule, we have

$$\begin{aligned}
\frac{\partial \ell_{ci}(\boldsymbol{\theta})}{\partial \alpha_j} &= \frac{\partial \boldsymbol{\Psi}}{\partial \alpha_j}\frac{\partial \ell_{ci}(\boldsymbol{\theta})}{\partial \boldsymbol{\Psi}} \\
&= \frac{\partial \boldsymbol{\Psi}}{\partial \alpha_j}\frac{1}{2}\text{tr}\{\boldsymbol{\Psi}^{-1}(\mathbf{b}_i\mathbf{b}_i^\top - \boldsymbol{\Psi})\boldsymbol{\Psi}^{-1}\}.
\end{aligned} \tag{C.2}$$

where, using the fact that $\text{tr}\{\mathbf{ABCD}\} = (\text{vec}(\mathbf{A}^\top))^\top(\mathbf{D}^\top \otimes \mathbf{B})(\text{vec}(\mathbf{C}))$, (C.2) can be rewritten as

$$\frac{\partial \ell_{ci}(\boldsymbol{\theta})}{\partial \alpha_j} = (\text{vec}(\frac{\partial \boldsymbol{\Psi}^\top}{\partial \alpha_j}))^\top \frac{1}{2}(\boldsymbol{\Psi}^{-1} \otimes \boldsymbol{\Psi}^{-1})(\text{vec}(\mathbf{b}_i\mathbf{b}_i^\top - \boldsymbol{\Psi})). \tag{C.3}$$

Let $\mathscr{D}_q$ be the elimination matrix (Lavielle, 2014) that transforms the vectorized $\boldsymbol{\Psi}$ (written as $\text{vec}(\boldsymbol{\Psi})$) into its half-vectorized form $\text{vech}(\boldsymbol{\Psi})$, such that $\mathscr{D}_q\text{vec}(\boldsymbol{\Psi}) = \text{vech}(\boldsymbol{\Psi})$. Using the fact that for all $j = 1, \ldots, \frac{1}{2}q(q+1)$, the vector $(\text{vec}(\frac{\partial \boldsymbol{\Psi}}{\partial \alpha_j})^\top)^\top$ corresponds to the jth row of the elimination matrix $\mathscr{D}_q$, we can generalize the derivative in (C.3) for the vector of parameters $\boldsymbol{\alpha}$ as

$$\frac{\partial \ell_{ci}(\boldsymbol{\theta})}{\partial \boldsymbol{\alpha}} = \frac{1}{2}\mathscr{D}_q(\boldsymbol{\Psi}^{-1} \otimes \boldsymbol{\Psi}^{-1})(\text{vec}(\mathbf{b}_i\mathbf{b}_i^\top - \boldsymbol{\Psi})).$$

Finally, at each iteration, we can compute the empirical information matrix by approximating the score for the observed log-likelihood by the stochastic approximation given in (21).

Appendix D Figures

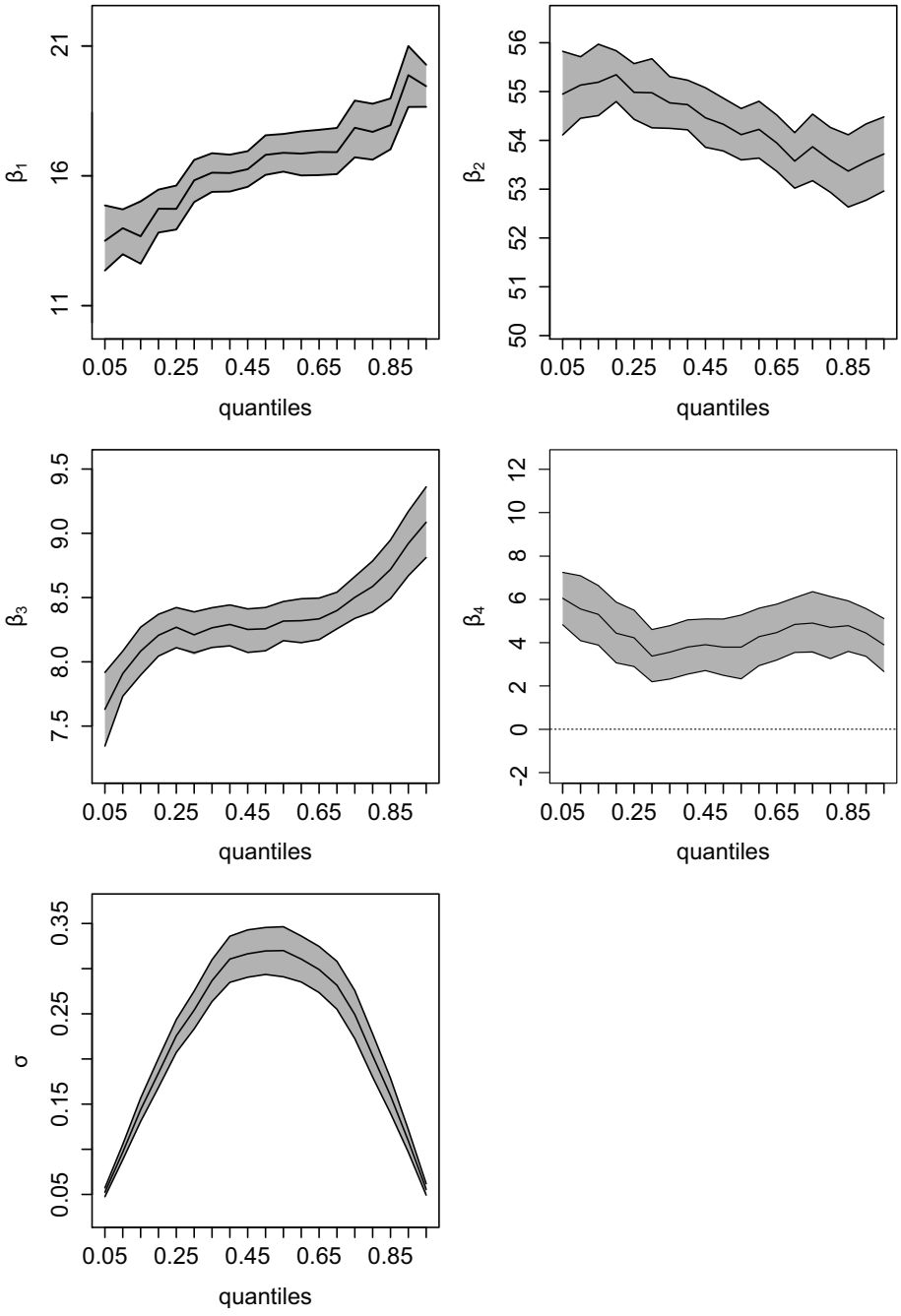

Figure 9: Soybean data: Point estimates (center solid line) and 95% confidence intervals for model parameters after fitting the QR-NLME model. The interpolated curves are spline-smoothed.

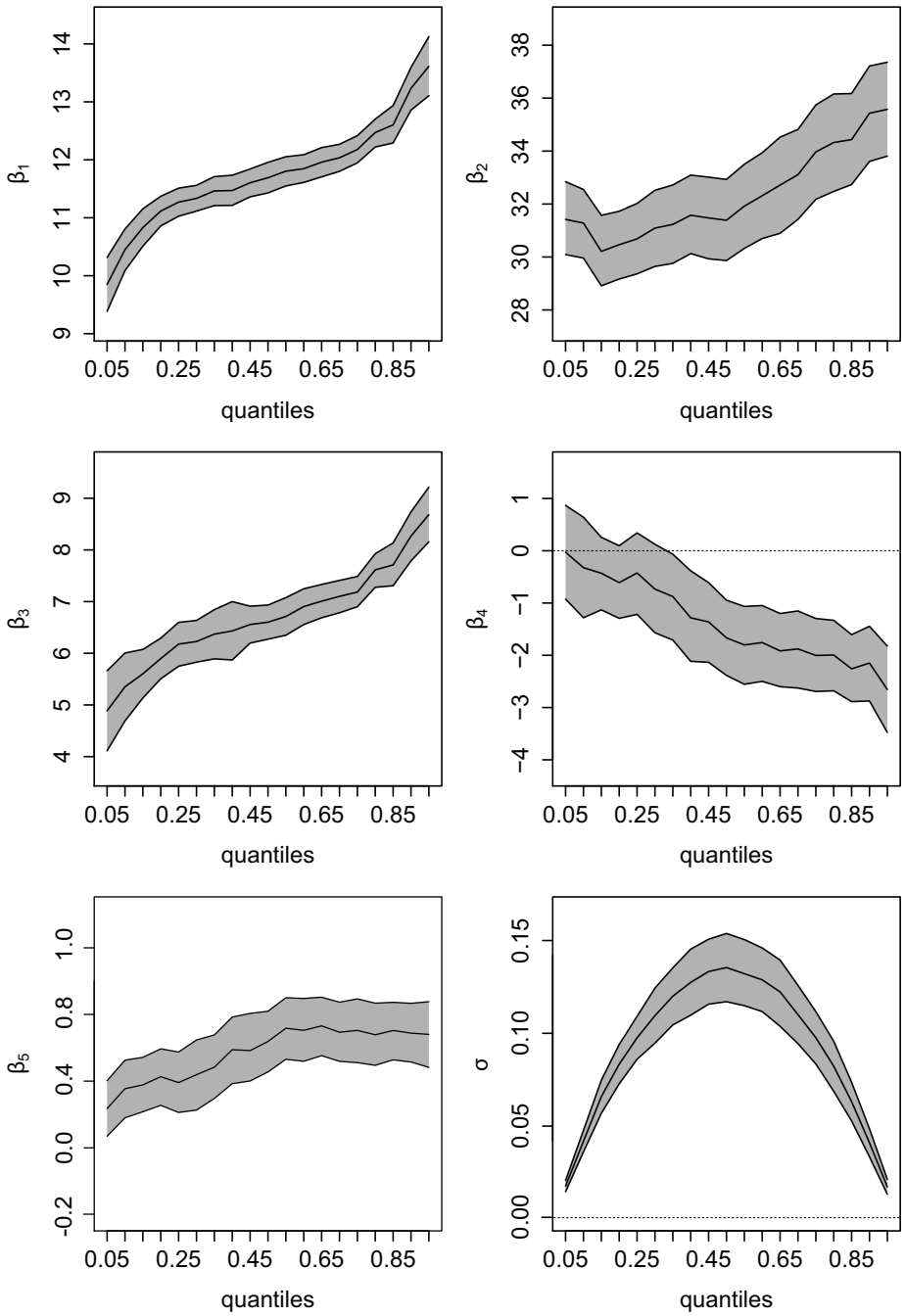

Figure 10: ACTG 315 data: Point estimates (center solid line) and 95% confidence intervals for model parameters after fitting the QR-NLME model. The interpolated curves are spline-smoothed.

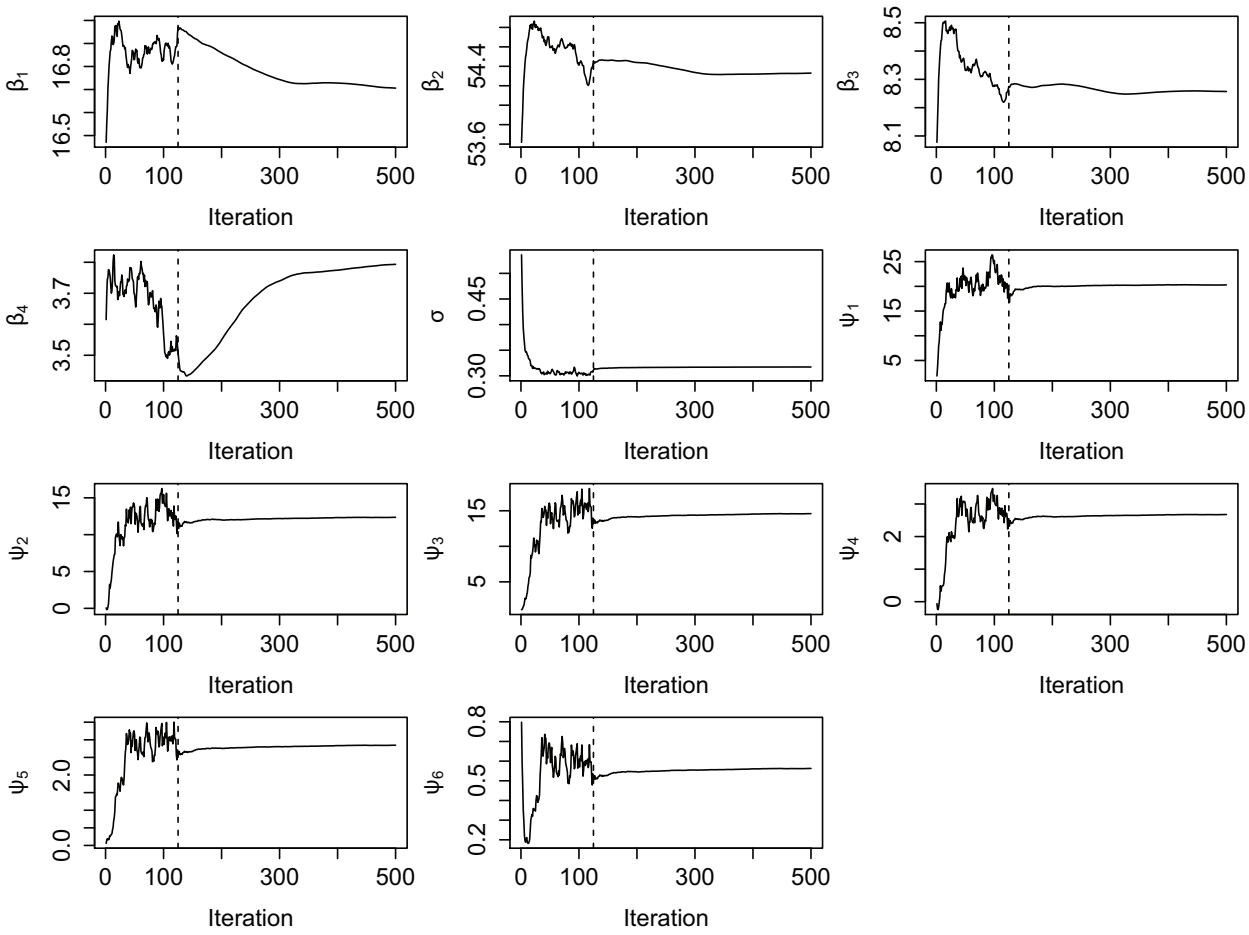

Figure 11: Graphical summary for the convergence of the fixed effect estimates, variance components of the random effects, and nuisance parameters performing a median regression ($p = 0.50$) for the Soybean data. The vertical dashed line delimits the beginning of the almost sure convergence as defined by the cut-point parameter $c = 0.25$.

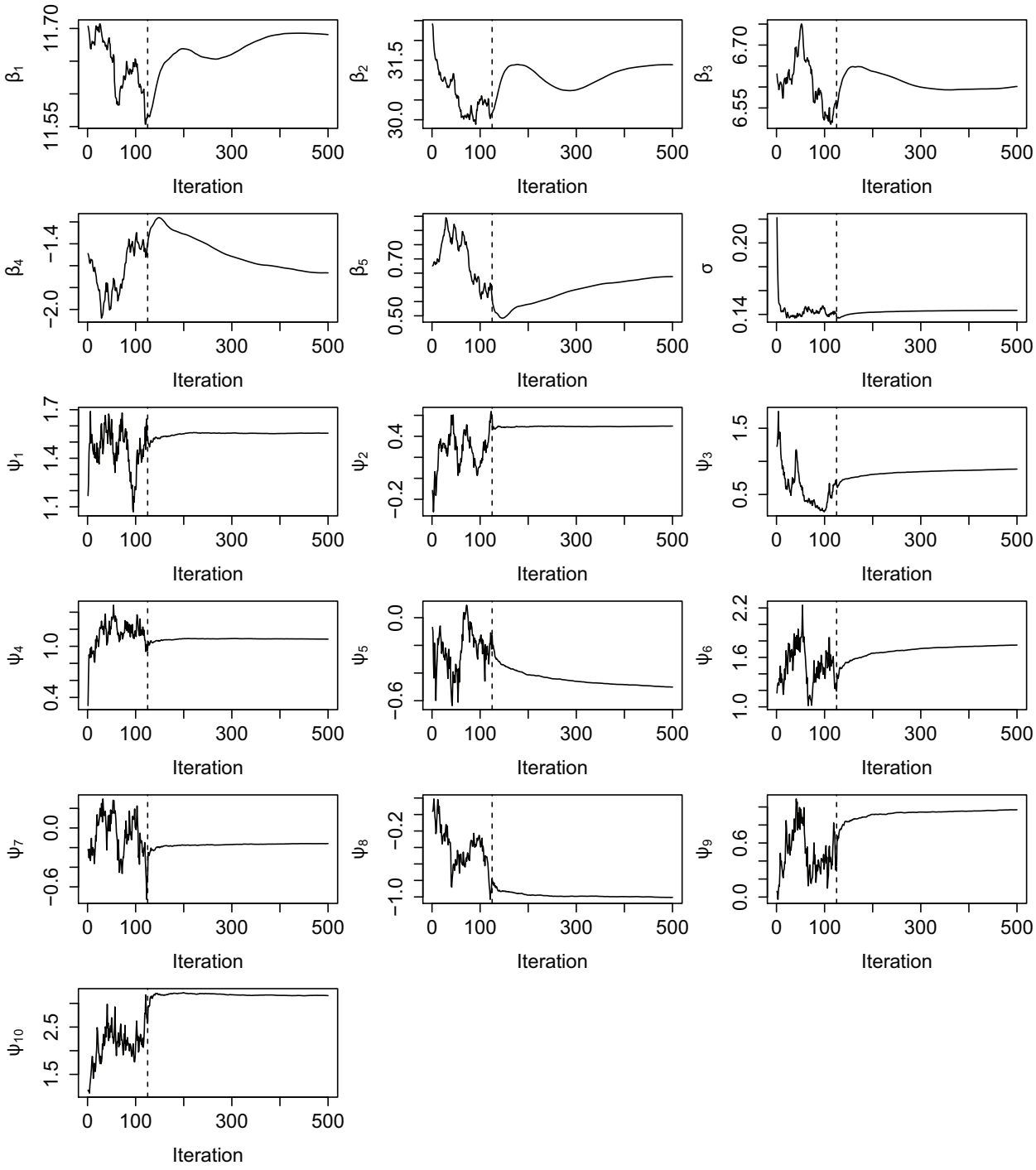

Figure 12: Graphical summary for the convergence of the fixed effect estimates, variance components of the random effects, and nuisance parameters performing a median regression ($p = 0.50$) for the HIV data. The vertical dashed line delimits the beginning of the almost sure convergence as defined by the cut-point parameter $c = 0.25$.

Appendix E Sample output from R package qrNLMM()

```
-------------------------------------------------
Quantile Regression for Nonlinear Mixed Model
-------------------------------------------------
Quantile = 0.5
Subjects = 48 ; Observations = 412

- Nonlinear function

function(x,fixed,random,covar=NA){
  resp = (fixed[1] + random[1])/(1 + exp(((fixed[2] +
          random[2]) - x)/(fixed[3] + random[3])))
  return(resp)}

-----------
Estimates
-----------
- Fixed effects

       Estimate Std. Error   z value Pr(>|z|)
beta 1 18.80029    0.53098  35.40704        0
beta 2 54.47930    0.29571 184.23015        0
beta 3  8.25797    0.09198  89.78489        0

sigma = 0.31569

Random effects Variance-Covariance Matrix matrix
         b1       b2      b3
b1 24.36687 12.27297 3.24721
b2 12.27297 15.15890 3.09129
b3  3.24721  3.09129 0.67193

------------------------
Model selection criteria
------------------------
        Loglik      AIC      BIC       HQ
Value -622.899 1265.798 1306.008 1281.703

-------
Details
-------
Convergence reached? = FALSE
Iterations = 300 / 300
Criteria = 0.00058
MC sample = 20
Cut point = 0.25
Processing time = 22.83885 mins
```

References

Allassonnière, S., Kuhn, E., Trouvé, A. et al. (2010). Construction of Bayesian deformable models via a stochastic approximation algorithm: a Convergence study. *Bernoulli*, **16**(3), 641–678.

Barndorff-Nielsen, O. E. & Shephard, N. (2001). Non-gaussian ornstein–uhlenbeck-based models and some of their uses in financial economics. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **63**(2), 167–241.

Booth, J. G. & Hobert, J. P. (1999). Maximizing generalized linear mixed model likelihoods with an automated monte carlo em algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **61**(1), 265–285.

Davidian, M. & Giltinan, D. (2003). Nonlinear models for repeated measurement data: an overview and update. *Journal of Agricultural, Biological and Environmental Statistics*, **8**(4), 387–419.

Davidian, M. & Giltinan, D. M. (1995). *Nonlinear Models for Repeated Measurement Data*, volume 62. CRC Press.

Delyon, B., Lavielle, M. & Moulines, E. (1999). Convergence of a stochastic approximation version of the EM algorithm. *Annals of Statistics*, **8**, 94–128.

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

Fu, L. & Wang, Y.-G. (2012). Quantile regression for longitudinal data with a working correlation model. *Computational Statistics & Data Analysis*, **56**(8), 2526–2538.

Galarza, C.E. and Benites, L.E. and Lachos, Victor V.H. (2015). *lqr: Robust Linear Quantile Regression*. R Foundation for Statistical Computing. package version 1.1.

Galvao, A. F. & Montes-Rojas, G. V. (2010). Penalized quantile regression for dynamic panel data. *Journal of Statistical Planning and Inference*, **140**(11), 3476–3497.

Galvao Jr, A. F. (2011). Quantile regression for dynamic panel data with fixed effects. *Journal of Econometrics*, **164**(1), 142–157.

Geraci, M. & Bottai, M. (2007). Quantile regression for longitudinal data using the asymmetric Laplace distribution. *Biostatistics*, **8**(1), 140–154.

Geraci, M. & Bottai, M. (2014). Linear quantile mixed models. *Statistics and Computing*, **24**(3), 461–479.

Grossman, Z., Polis, M., Feinberg, M. B., Grossman, Z., Levi, I., Jankelevich, S., Yarchoan, R., Boon, J., de Wolf, F., Lange, J. M. et al. (1999). Ongoing hiv dissemination during haart. *Nature medicine*, **5**(10), 1099–1104.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**(1), 97–109.

Huang, Y. & Dagne, G. (2011). A bayesian approach to joint mixed-effects models with a skew-normal distribution and measurement errors in covariates. *Biometrics*, **67**(1), 260–269.

Koenker, R. (2004). Quantile regression for longitudinal data. *Journal of Multivariate Analysis*, **91**(1), 74–89.

Koenker, R. (2005). *Quantile Regression*. Cambridge University Press, New York, NY.

Koenker, R. & Machado, J. (1999). Goodness of fit and related inference processes for quantile regression. *Journal of the American Statistical Association*, **94**(448), 1296–1310.

Kotz, S., Kozubowski, T. & Podgorski, K. (2001). *The Laplace distribution and generalizations: A revisit with applications to communications, economics, engineering, and finance*. Birkhauser.

Kuhn, E. & Lavielle, M. (2004). Coupling a stochastic approximation version of EM with an MCMC procedure. *ESAIM: Probability and Statistics*, **8**, 115–131.

Kuhn, E. & Lavielle, M. (2005). Maximum likelihood estimation in nonlinear mixed effects models. *Computational Statistics & Data Analysis*, **49**(4), 1020–1038.

Kuzobowski, T. J. & Podgorski, K. (2000). A multivariate and asymmetric generalization of laplace distribution. *Computational Statistics*, **15(4)**, 531–540.

Lachos, V. H., Ghosh, P. & Arellano-Valle, R. B. (2010). Likelihood based Inference for Skew–Normal Independent Linear Mixed Models. *Statistica Sinica*, **20**(1), 303–322.

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

Lavielle, M. (2014). *Mixed Effects Models for the Population Approach*. Chapman and Hall/CRC, Boca Raton, FL.

Lipsitz, S. R., Fitzmaurice, G. M., Molenberghs, G. & Zhao, L. P. (1997). Quantile Regression Methods for Longitudinal Data with Drop-outs: Application to CD4 Cell Counts of Patients Infected with the Human Immunodeficiency Virus. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **46**(4), 463–476.

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society - Series B (Methodological)*, **44**(2), 226–233.

Meilijson, I. (1989). A fast improvement to the EM algorithm on its own terms. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 127–138.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. (1953). Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, **21**, 1087–1092.

Meza, C., Osorio, F. & De la Cruz, R. (2012). Estimation in nonlinear mixed-effects models using heavy-tailed distributions. *Statistics and Computing*, **22**, 121–139.

Perelson, A. S., Essunger, P., Cao, Y., Vesanen, M., Hurley, A., Saksela, K., Markowitz, M. & Ho, D. D. (1997). Decay characteristics of hiv-1-infected compartments during combination therapy.

Pinheiro, J. & Bates, D. (1995). Approximations to the log-likelihood function in the nonlinear mixed effects model. *Journal of Computational and Graphical Statistics*, **4**, 12–35.

Pinheiro, J. C. & Bates, D. M. (2000). *Mixed-effects Models in S and S-PLUS*. Springer, New York, NY.

Searle, S. R., Casella, G. & McCulloch, C. (1992). Variance components, 1992.

Vaida, F. (2005). Parameter convergence for EM and MM algorithms. *Statistica Sinica*, **15**(3), 831–840.

Wang, J. (2012). Bayesian quantile regression for parametric nonlinear mixed effects models. *Statistical Methods and Applications*, **21**, 279–295.

Wei, G. C. & Tanner, M. A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association*, **85**(411), 699–704.

Wichitaksorn, N., Choy, S. & Gerlach, R. (2014). A generalized class of skew distributions and associated robust quantile regression models. *Canadian Journal of Statistics*, **42**(4), 579–596.

Wu, C. J. (1983). On the convergence properties of the em algorithm. *The Annals of Atatistics*, pages 95–103.

Wu, L. (2002). A joint model for nonlinear mixed-effects models with censoring and covariates measured with error, with application to aids studies. *Journal of the American Statistical Association*, **97**(460), 955–964.

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

Yu, K. & Moyeed, R. (2001). Bayesian quantile regression. *Statistics & Probability Letters*, **54**(4), 437–447.

Yu, K. & Zhang, J. (2005). A three-parameter asymmetric Laplace distribution and its extension. *Communications in Statistics - Theory and Methods*, **34**(9-10), 1867–1879.

Yuan, Y. & Yin, G. (2010). Bayesian quantile regression for longitudinal studies with nonignorable missing data. *Biometrics*, **66**(1), 105–114.