LIKELIHOOD BASED INFERENCE FOR QUANTILE REGRESSION IN NONLINEAR MIXED EFFECTS MODELS

CHRISTIAN E. GALARZA
Escuela Superior Politécnica del Litoral, Guayaquil, Ecuador.
cgalarza88@gmail.com; (+593) 4 2210505

VICTOR H. LACHOS
Departamento de Estatística, IMECC, Universidade Estadual de Campinas, Campinas, São Paulo, Brazil.
hlachos@ime.unicamp.br; (+55) 19 35216078

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 for analyzing QR models for correlated continuous longitudinal data via the asymmetric Laplace distribution (ALD). Exploiting the nice hierarchical representation of the ALD, 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 (NLMMs). 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.

RESUMEN

Los datos longitudinales son frecuentemente analizados usando modelos de efectos mixtos normales. Por otra parte, los métodos de estimación tradicionales son basados en regresión en media, lo cual conduce a estimaciones no robustas de los parámetros cuando los errores no se distribuyen normalmente. Comparada con el enfoque de la regresión en media tradicional, la regresión cuantílica (RC) puede caracterizar completamente la distribución condicional de la variable de respuesta y es más robusta ante la presencia de valores atípicos y especificaciones erróneas de la distribución del error. Este artículo usa un enfoque basado en verosimilitud para analizar modelos de RC para datos continuos longitudinales correlacionados usando la distribución Laplace asimétrica (DLA). Haciendo uso de la representación estocástica de la DLA, nuestro enfoque clásico utiliza una Aproximación Estocástica del algoritmo EM (SAEM) para conseguir estimativas de máxima verosimilitud

(MV) exactas para los efectos fijos y los componentes de varianza en modelos no lineales de efectos mixtos. Evaluamos el desempeño del algoritmo en muestras finitas y las propiedades asintóticas de las estimativas de MV a través de experimentos empíricos y aplicaciones para dos conjuntos de datos reales. El algoritmo SAEM propuesto se encuentra implementado en el paquete de R `qrNLMM`.

Palabras clave

Distribución Laplace asimétrica, Modelos no lineales de Efectos Mixtos, Regresión cuantílica, Algoritmo SAEM, Aproximaciones Estocásticas.

1 Introduction

Nonlinear mixed-effects models (NLMMs) are frequently used to analyze grouped, clustered, longitudinal and multilevel data because of their potential to handle, on one hand, nonlinearities in the relationship between the observed response and the covariates and random effects, and on the other hand, to take into account within and between-subject correlations presented in this type of data (Pinheiro & Bates, 2000; Davidian & Giltinan, 2003; Wu, 2010). Moreover, NLMMs are also flexible and often mechanistic, based on biological, chemical, physics mechanisms, among others, leading to a natural modelling using a known family of nonlinear functions providing desirable characteristics such as asymptotes, a unique maximum value, monotonicity, positive range, etc. Majority of these NLMMs estimate covariate effects on the response through a mean regression, controlling for between-cluster heterogeneity via normally-distributed cluster-specific random effects and random errors. However, this centrality-based inferential framework is often inadequate when the conditional distribution of the response (conditional on the random terms) is skewed, multimodal, or affected by atypical observations. In contrast, conditional quantile regression (QR) methods (Koenker, 2004, 2005) quantifying the entire conditional distribution of the outcome variable were developed that can provide assessment of covariate effects at any arbitrary quantiles of the outcome. In addition, QR methods do not impose any distribution assumption on the error, except requiring that the error term has a zero conditional quantile such as the ALD. Because of its popularity and the flexibility it provides, standard QR methods are implementable via available software packages, for example, the R package `quantreg()`.

Although QR was initially developed under a univariate framework, the abundance of clustered data in recent times lead to its extensions into mixed modeling framework via either the distribution-free route Lipsitz *et al.* (1997); Galvao & Montes-Rojas (2010); Galvao Jr (2011); Fu & Wang (2012), or the traditional likelihood-based route mostly using the ALD Geraci & Bottai (2007); Yuan & Yin (2010); Geraci & Bottai (2014). Among the ALD-based models, Geraci & Bottai (2007) proposed a Monte Carlo EM (MCEM)-based conditional QR model for continuous responses with a subject-specific random (univariate) intercept to account for within-subject dependence in the context of longitudinal data. However, due to the limitations of a simple random intercept model to account for the between-cluster heterogeneity, Geraci & Bottai (2014) extended their previous Geraci & Bottai (2007) model to a general linear quantile mixed effects regression model (QR-LMM) with multiple random effects (both intercepts and slopes). However, instead of going the MCEM route, the estimation of the fixed effects and the covariance components were implemented

using an efficient combination of Gaussian quadrature approximations and non-smooth optimization algorithms. Yuan & Yin (2010) applied the version of QR of Geraci & Bottai (2007) to linear mixed effects models for longitudinal measurements with missing data. Wang (2012) considered QR-NLMMs from a Bayesian perspective and shown that QR-NLMMs may be a better measure of centrality for skewed or multimodal data and more robust against nonnormality of the distribution of random errors than the mean regression estimator. Although some results on QR-NLMMs have recently appeared in the literature, to the best of our knowledge, there seem to be no studies on exact inference for QR-NLMMs from a likelihood based perspective.

In this paper, we proceed to achieve that via a robust parametric ALD-based QR-NLMMs, where the full likelihood-based implementation follows a stochastic version of the EM algorithm (SAEM), proposed by Delyon *et al.* (1999), for maximum likelihood (ML) estimation in contrast to the approximations proposed by Geraci & Bottai (2014) for QR-LMMs. The SAEM algorithm has been proved to be more computationally efficient than the classical MCEM algorithm due to the recycling of simulations from one iteration to the next in the smoothing phase of the algorithm. 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 in computing the ML estimates in mixed effects models. Our empirical results shows that the ML estimates based on the SAEM algorithm do provide good asymptotic properties. Furthermore, application of our method to two longitudinal datasets is illustrated via the `R` package `qrNLMM()`.

The rest of the paper proceeds as follows. Section 2 presents some preliminaries, in particular the connection between QR and ALD and an outline of the EM and SAEM algorithms. Section 3 develops the MCEM and the SAEM algorithms for a general NLMM, while Section 4 outlines the likelihood estimation and standard errors. Section 5 presents some simulation studies. Application of the SAEM method to two longitudinal datasets, one examining the Soybean genotypes data and the other on a HIV viral load study are presented in Section 6. Finally, Section 7 concludes, sketching some future research directions.

2 Preliminaries

In this section, we provide some useful results on the ALD and QR, and introduce the EM and SAEM algorithms for ML estimation.

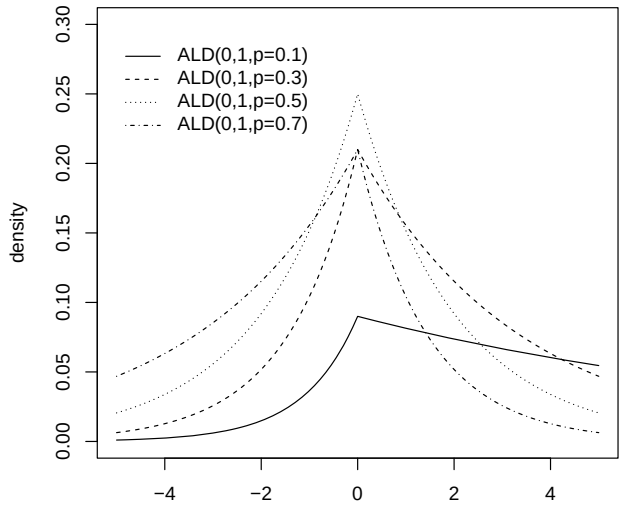
2.1 Connection between QR and ALD

Following Yu & Moyeed (2001), a random variable Y is distributed as an ALD with location parameter μ, scale parameter $\sigma > 0$ and skewness parameter $p \in (0,1)$, if its probability density function (pdf) is given by

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

where $\rho_p(.)$ is the check (or loss) function defined by $\rho_p(u) = u(p - \mathbb{I}\{u < 0\})$, with $\mathbb{I}\{.\}$ the usual indicator function. This distribution is denoted by $ALD(\mu, \sigma, p)$. It is easy to see that $W = \rho_p\left(\frac{Y-\mu}{\sigma}\right)$ follows an exponential(1) distribution. Figure 1 plots the ALD illustrating how the the skewness changes with altering choices for 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 ALD have equal mass and the distribution resemble the more common double exponential distribution. In contrast to the normal distribution with a quadratic term in the exponent, the ALD is linear in the exponent. This results in a more peaked mode for the ALD together with thicker tails. On the contrary, the normal distribution has heavier shoulders compared to the ALD.

Figure 1. Standard asymmetric Laplace density



ALD abides by the following stochastic representation (Kotz *et al.*, 2001; Kuzobowski & Podgorski, 2000). Let $U \sim \exp(\sigma)$ and $Z \sim N(0,1)$ be two independent random variables. Then, $Y \sim ALD(\mu, \sigma, p)$ can be represented as

$$Y \stackrel{d}{=} \mu + \vartheta_p U + \tau_p \sqrt{\sigma U} Z, \tag{2}$$

where $\vartheta_p = \frac{1-2p}{p(1-p)}$ and $\tau_p^2 = \frac{2}{p(1-p)}$, and $\stackrel{d}{=}$ denotes equality in distribution. This representation is useful in obtaining the moment generating function (mgf), and formulating the estimation algorithm. From (2), the hierarchical representation of the ALD is given as

$$\begin{aligned} Y|U = u &\sim N(\mu + \vartheta_p u, \tau_p^2 \sigma u), \\ U &\sim \exp(\sigma). \end{aligned} \tag{3}$$

This representation will be useful for the implementation of the EM algorithm. Moreover, since $Y|U = u \sim N(\mu + \vartheta_p u, \tau_p^2 \sigma u)$, one can easily derive the pdf of Y, given by

$$f(y|\mu, \sigma, p) = \frac{1}{\sqrt{2\pi}} \frac{1}{\tau_p \sigma^{\frac{3}{2}}} \exp\left(\frac{\delta(y)}{\gamma}\right) A(y), \tag{4}$$

where $\delta(y) = \frac{|y-\mu|}{\tau_p\sqrt{\sigma}}$, $\gamma = \sqrt{\frac{1}{\sigma}\left(2+\frac{\vartheta_p^2}{\tau_p^2}\right)} = \frac{\tau_p}{2\sqrt{\sigma}}$ and $A(y) = 2\left(\frac{\delta(y)}{\gamma}\right)^{1/2} K_{1/2}\big(\delta(y)\gamma\big)$, with $K_\nu(.)$, the modified Bessel function of the third kind. It easy to see that that the conditional distribution of U, given $Y = y$, is $U|(Y = y) \sim GIG(\frac{1}{2}, \delta, \gamma)$, where $GIG(\nu, a, b)$ represents the Generalized Inverse Gaussian (GIG) distribution (Barndorff-Nielsen & Shephard, 2001) with the pdf

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

The moments of U can be expressed as

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

Some useful properties of the Bessel function of the third kind $K_\lambda(u)$ are: (i) $K_\nu(u) = K_{-\nu}(u)$; (ii) $K_{\nu+1}(u) = \frac{2\nu}{u}K_\nu(u) + K_{\nu-1}(u)$; (iii) for non-negative integer r, $K_{r+1/2}(u) = \sqrt{\frac{\pi}{2u}}\exp(-u)$ $\sum_{k=0}^{r} \frac{(r+k)!(2u)^{-k}}{(r-k)!k!}$. A special case is $K_{1/2}(u) = \sqrt{\frac{\pi}{2u}}\exp(-u)$.

2.2 The EM and SAEM algorithms

In models with missing data, the EM algorithm (Dempster *et al.*, 1977) has established itself as the most popular tool for obtaining the ML estimates of the model parameters. This iterative algorithm maximizes the complete log-likelihood function $\ell_c(\boldsymbol{\theta}; \mathbf{y}_{\text{com}})$ at each step, converging quickly 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 $Q(\boldsymbol{\theta}|\widehat{\theta}^{(k)}) = \mathrm{E}\left\{\ell_c(\boldsymbol{\theta}; \mathbf{y}_{\text{com}})|\widehat{\boldsymbol{\theta}}^{(k)}, \mathbf{y}_{\text{obs}}\right\}$, where $\widehat{\boldsymbol{\theta}}^{(k)}$ is the estimate of $\boldsymbol{\theta}$ at the k-th iteration;

M-Step: Maximize $Q(\theta|\widehat{\boldsymbol{\theta}}^{(k)})$ with respect to $\boldsymbol{\theta}$ obtaining $\widehat{\boldsymbol{\theta}}^{(k+1)}$.

However, in some applications of the EM algorithm, the E-step cannot be obtained analytically and has to be calculated using simulations. 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. This simple solution is infact computationally expensive, given the need to generate a large number of independent simulations of the missing data for a good approximation. Thus, in order to reduce the amount of required simulations compared to the MCEM algorithm, the SAEM algorithm proposed by Delyon *et al.* (1999) replaces the E-step of the EM algorithm by a stochastic approximation procedure, while the Maximization step remains unchanged. Besides having good theoretical properties, the SAEM 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 with the conditional distribution, and updates the unknown parameters of the model. 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 $f(\mathbf{q}|\theta^{(k-1)},\mathbf{y}_i)$.
- Stochastic Approximation: Update the $Q(\theta|\widehat{\theta}^{(k)})$ function as

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

M-Step:

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

where δ_k is a smoothness parameter (Kuhn & Lavielle, 2004), i.e., a decreasing sequence of positive numbers such that $\sum_{k=1}^{\infty}\delta_k = \infty$ and $\sum_{k=1}^{\infty}\delta_k^2 < \infty$. Note that, for the SAEM algorithm, the E-Step coincides with the MCEM algorithm, however a small number of simulations m (suggested to be $m \leq 20$) is necessary. 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 at the iteration k denoted by $(\mathbf{q}^{(\ell,k)})$, $\ell = 1,\ldots,m$ but some or all previous simulations, where this 'memory' property is set by the smoothing parameter δ_k.

Note, in equation (6), if the smoothing parameter δ_k is equal to 1 for all k, the SAEM algorithm will have 'no memory', and will be equivalent to the MCEM algorithm. The SAEM with no memory will converge quickly (convergence in distribution) to a solution neighbourhood, however when the algorithm with memory will converge slowly (almost sure convergence) to the ML solution. We suggested the following choice of the smoothing parameter given as

$$\delta_k = \begin{cases} 1, & \text{for} \quad 1 \leq k \leq cW \\ \frac{1}{k-cW}, & \text{for} \quad cW+1 \leq k \leq W \end{cases} \tag{7}$$

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. For example, if $c = 0$ the algorithm will have memory for all iterations, and hence will converge slowly to the ML estimates. If $c = 1$, the algorithm will have no memory, and so will converge quickly to a solution neighbourhood. For the first case, W would need to be large in order to achieve the ML estimates. For the second, the algorithm will output a Markov Chain where after applying a *burn in* and *thin*, the mean of the chain observations can be a reasonable estimate.

A number between 0 and 1 ($0 < c < 1$) will assure an initial convergence in distribution to a solution neighbourhood for the first cW iterations and an almost sure convergence for the rest of the iterations. Hence, this combination will leads us to a fast algorithm with good estimates.

To implement SAEM, the user must fix several constants matching the number of total iterations W and the cut point c that defines the starting of the smoothing step of the SAEM algorithm, however those parameters will vary depending of the model and the data. To determinate those constants, a graphical approach is recommended to monitor the convergence of the estimates for all the parameters, and, if possible, to monitor the difference (relative difference) between two successive evaluations of the log-likelihood $\ell(\boldsymbol{\theta}|\mathbf{y}_{obs})$, given by $||\ell(\boldsymbol{\theta}^{(k+1)}|\mathbf{y}_{obs}) - \ell(\boldsymbol{\theta}^{(k)}|\mathbf{y}_{obs})||$ or $||\ell(\boldsymbol{\theta}^{(k+1)}|\mathbf{y}_{obs})/\ell(\boldsymbol{\theta}^{(k)}|\mathbf{y}_{obs}) - 1||$, respectively.

3 QR for nonlinear mixed models and algorithms

We proposed the following general mixed-effects model. Let $\mathbf{y}_i = (y_{i1}, ..., y_{in_i})^\top$ denote the continuous response for subject i and let $\eta = (\eta(\phi_i, x_{i1}), ..., \eta(\phi_i, x_{in_i}))^\top$ represents a nonlinear differentiable function of vector-valued mixed-effects random parameters ϕ_i of dimension r and a matrix of covariates $\mathbf{x}_i$ of dimensions $n_i \times r$. We define the NLMM as

$$\mathbf{y}_i = \eta(\boldsymbol{\phi}_i, \mathbf{x}_i) + \boldsymbol{\varepsilon}_i, \qquad \phi_i = \mathbf{A}_i\boldsymbol{\beta}_p + \mathbf{B}_i\mathbf{b}_i, \tag{8}$$

where $\mathbf{A}_i$ and $\mathbf{B}_i$ are design matrices 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 and ε_i the independent and identically distributed vector of random errors. We define pth quantile function of the response y_{ij} as

$$Q_p(y_{ij}|\mathbf{x}_{ij}, \mathbf{b}_i) = \eta(\phi_i, x_{ij}) = \eta(\mathbf{A}_i\boldsymbol{\beta}_p + \mathbf{B}_i\mathbf{b}_i, x_{ij}). \tag{9}$$

where Q_p denotes the inverse of the unknown distribution function F, the random effects $\mathbf{b}_i$ are distributed as $\mathbf{b}_i \overset{\text{iid}}{\sim} \mathrm{N}_q(\mathbf{0}, \boldsymbol{\Psi})$, where the dispersion matrix $\boldsymbol{\Psi} = \boldsymbol{\Psi}(\boldsymbol{\alpha})$ depends on unknown and reduced parameters $\boldsymbol{\alpha}$, and the errors are distributed as $\varepsilon_{ij} \overset{\text{iid}}{\sim} ALD(0, \sigma)$ and both uncorrelated. Then, $y_{ij}|\mathbf{b}_i$ independently follows as ALD with the density given by

$$f(y_{ij}|\boldsymbol{\beta}_p, \mathbf{b}_i, \sigma) = \frac{p(1-p)}{\sigma} \exp\left\{-\rho_p\left(\frac{y_{ij} - \eta(\mathbf{A}_i\boldsymbol{\beta}_p + \mathbf{B}_i\mathbf{b}_i, \mathbf{x}_{ij})}{\sigma}\right)\right\}. \tag{10}$$

3.1 A MCEM algorithm

First, we develop a MCEM algorithm for ML estimation of the parameters in the QR-NLMM. The model exhibits a flexible hierarchical representation, which is useful in deriving the theoretical properties. From (3), the QR-NLMM defined in (9)-(10), can be represented in a hierarchical form as:

$$\begin{aligned} \mathbf{y}_i|\mathbf{b}_i, \mathbf{u}_i &\sim N_{n_i}\left(\boldsymbol{\eta}(\mathbf{A}_i\boldsymbol{\beta}_p + \mathbf{B}_i\mathbf{b}_i, \mathbf{x}_i) + \vartheta_p\mathbf{u}_i, \sigma\tau_p^2\mathbf{D}_i\right), \\ \mathbf{b}_i &\sim N_q(\mathbf{0}, \boldsymbol{\Psi}), \\ \mathbf{u}_i &\sim \prod_{j=1}^{n_i} \exp(\sigma), \end{aligned} \tag{11}$$

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

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

where

$$\begin{aligned}\ell_c(\boldsymbol{\theta}; \mathbf{y}_{i_c}) = \text{constant} &- \frac{3}{2} n_i log\sigma - \frac{1}{2} log\left|\boldsymbol{\Psi}\right| - \frac{1}{2} \mathbf{b}_i^\top \boldsymbol{\Psi}^{-1} \mathbf{b}_i - \frac{1}{\sigma} \mathbf{u}_i^\top \mathbf{1}_{n_i} \\ &- \frac{1}{2\sigma\tau_p^2} (\mathbf{y}_i - \boldsymbol{\eta}(\mathbf{A}_i\boldsymbol{\beta}_p + \mathbf{B}_i\mathbf{b}_i, \mathbf{x}_i) - \vartheta_p \mathbf{u}_i)^\top \mathbf{D}_i^{-1} (\mathbf{y}_i - \boldsymbol{\eta}(\mathbf{A}_i\boldsymbol{\beta}_p + \mathbf{B}_i\mathbf{b}_i, \mathbf{x}_i) - \vartheta_p \mathbf{u}_i). \quad (12)\end{aligned}$$

Since $\mathbf{A}_i$, $\mathbf{B}_i$ and $\mathbf{x}_i$ are known matrices, we will 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

$$Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k)}) = \textstyle\sum_{i=1}^{n} Q_i(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k)}),$$

where

$$\begin{aligned}Q_i(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k)}) &= \mathrm{E}\left\{\ell_c(\boldsymbol{\theta}; \mathbf{y}_{i_c})|\boldsymbol{\theta}^{(k)}, \mathbf{y}\right\} \quad (13) \\ &\propto -\frac{3}{2} n_i log\sigma - \frac{1}{2} log\left|\boldsymbol{\Psi}\right| - \frac{1}{2} \mathrm{tr}\left\{\widehat{(\mathbf{b}\mathbf{b}^\top)_i}^{(k)} \boldsymbol{\Psi}^{-1}\right\} - \frac{1}{2\sigma\tau_p^2}\left[\mathbf{y}_i^\top \widehat{\mathbf{D}_i^{-1}}^{(k)} \mathbf{y}_i\right. \\ &\left. - 2\vartheta_p \mathbf{y}_i^\top \mathbf{1}_{n_i} + \frac{\tau_p^4}{4} \widehat{\mathbf{u}}_i^{(k)\top} \mathbf{1}_{n_i} - 2\mathbf{y}_i^\top \widehat{(\mathbf{D}^{-1}\boldsymbol{\eta})}_i^{(k)} + 2\vartheta_p \mathbf{1}_{n_i}^\top \widehat{\boldsymbol{\eta}}_i^{(k)} + \widehat{\boldsymbol{\eta}_i^\top \mathbf{D}_i^{-1}} \boldsymbol{\eta}_i^{(k)}\right]\end{aligned}$$

where $\boldsymbol{\eta}_i = \boldsymbol{\eta}(\mathbf{A}_i\boldsymbol{\beta}_p + \mathbf{B}_i\mathbf{b}_i, \mathbf{x}_i)$ for simplicity, $\mathrm{tr}(\mathbf{A})$ indicates the trace of matrix $\mathbf{A}$ and $\mathbf{1}_p$ is the vector of ones of dimension p. The calculation of these function requires expressions for

$$\begin{aligned}\widehat{\boldsymbol{\eta}}_i^{(k)} &= \mathrm{E}\left\{\boldsymbol{\eta}_i|\boldsymbol{\theta}^{(k)}, \mathbf{y}_i\right\}, & \widehat{\mathbf{u}}_i^{(k)} &= \mathrm{E}\left\{\mathbf{u}_i|\boldsymbol{\theta}^{(k)}, \mathbf{y}_i\right\}, \\ \widehat{(\mathbf{b}\mathbf{b}^\top)_i}^{(k)} &= \mathrm{E}\left\{\mathbf{b}_i\mathbf{b}_i^\top|\boldsymbol{\theta}^{(k)}, \mathbf{y}_i\right\}, & \widehat{\mathbf{D}_i^{-1}}^{(k)} &= \mathrm{E}\left\{\mathbf{D}_i^{-1}|\boldsymbol{\theta}^{(k)}, \mathbf{y}_i\right\}, \\ \widehat{(\mathbf{D}^{-1}\boldsymbol{\eta})}_i^{(k)} &= \mathrm{E}\left\{\mathbf{D}_i^{-1}\boldsymbol{\eta}_i^{(k)}|\boldsymbol{\theta}^{(k)}, \mathbf{y}_i\right\}, & \widehat{(\boldsymbol{\eta}^\top\mathbf{D}^{-1}\boldsymbol{\eta})}_i^{(k)} &= \mathrm{E}\left\{\boldsymbol{\eta}_i^\top \mathbf{D}_i^{-1} \boldsymbol{\eta}_i|\boldsymbol{\theta}^{(k)}, \mathbf{y}_i\right\},\end{aligned}$$

which do not have closed forms. Since the joint distribution of the missing data $(\mathbf{b}_i^{(k)}, \mathbf{u}_i^{(k)})$ is unknown and the conditional expectations cannot be computed analytically, for any function $g(.)$, the MCEM algorithm approximates the conditional expectations above by their Monte Carlo approximations

$$\mathrm{E}[g(\mathbf{b}_i, \mathbf{u}_i)|\boldsymbol{\theta}^{(k)}, \mathbf{y}_i] \approx \frac{1}{m} \sum_{\ell=1}^{m} g(\mathbf{b}_i^{(\ell,k)}, \mathbf{u}_i^{(\ell,k)}), \quad (14)$$

which depend of the simulations of the two latent (missing) variables $\mathbf{b}_i^{(k)}$ and $\mathbf{u}_i^{(k)}$ from the conditional joint density $f(\mathbf{b}_i,\mathbf{u}_i|\boldsymbol{\theta}^{(k)},\mathbf{y}_i)$. Using known properties of conditional expectations, the expected value in (14) can be more accurately approximated as

$$\begin{aligned}\mathrm{E}_{\mathbf{b}_i,\mathbf{u}_i}[g(\mathbf{b}_i,\mathbf{u}_i)|\boldsymbol{\theta}^{(k)},\mathbf{y}_i] &= \mathrm{E}_{\mathbf{b}_i}[\mathrm{E}_{\mathbf{u}_i}[g(\mathbf{b}_i,\mathbf{u}_i)|\boldsymbol{\theta}^{(k)},\mathbf{b}_i,\mathbf{y}_i]|\mathbf{y}_i] \\ &\approx \frac{1}{m}\sum_{\ell=1}^{m}\mathrm{E}_{\mathbf{u}_i}[g(\mathbf{b}_i^{(\ell,k)},\mathbf{u}_i)|\boldsymbol{\theta}^{(k)},\mathbf{b}_i^{(\ell,k)},\mathbf{y}_i], \end{aligned} \tag{15}$$

where $\mathbf{b}^{(\ell,k)}$ is a sample from the conditional density $f(\mathbf{b}_i|\boldsymbol{\theta}^{(k)},\mathbf{y}_i)$. Note that (15) is a more accurate approximation once it only depends of one MC approximation, instead two as needed in (14).

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

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

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

$$\chi_{ij}=\frac{|y_{ij}-\eta_{ij}(\boldsymbol{\beta}_p,\mathbf{b}_i)|}{\tau_p\sqrt{\sigma}} \quad \text{and} \quad \psi=\frac{\tau_p}{2\sqrt{\sigma}} \tag{16}$$

From (5), and after generating samples from $f(\mathbf{b}_i|\boldsymbol{\theta}^{(k)},\mathbf{y}_i)$ (see Subsection 4.2), the conditional expectation $\mathrm{E}_{\mathbf{u}_i}[\cdot|\boldsymbol{\theta},\mathbf{b}_i,\mathbf{y}_i]$ in (15) can be computed analytically. Finally, the proposed MCEM algorithm for estimating the parameters of the QR-NLMM can be summarized as follows:

MC E-step: Given $\boldsymbol{\theta}=\boldsymbol{\theta}^{(k)}$, for $i=1,\ldots,n$;

- **Simulation Step:** For $\ell=1,\ldots,m$, draw $\mathbf{b}_i^{(\ell,k)}$ from $f(\mathbf{b}_i|\boldsymbol{\theta}^{(k)},\mathbf{y}_i)$, as described later in Subsection 4.2.
- **Monte Carlo approximation:** Using (5) and the simulated sample above, evaluate

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

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

$$\begin{aligned}\widehat{\boldsymbol{\beta}_p}^{(k+1)} = \widehat{\boldsymbol{\beta}_p}^{(k)} &+ \left[\sum_{i=1}^{n}\left\{\frac{1}{m}\sum_{\ell=1}^{m}\mathbf{J}_i^{(k)^\top}\mathscr{E}(\mathbf{D}_i^{-1})^{(\ell,k)}\mathbf{J}_i^{(k)}\right\}\right]^{-1}\times \\ &\left[\sum_{i=1}^{n}\left\{\frac{1}{m}\sum_{\ell=1}^{m}\left[2\mathbf{J}_i^{(k)^\top}\mathscr{E}(\mathbf{D}_i^{-1})^{(\ell,k)}\left[\mathbf{y}_i-\boldsymbol{\eta}(\widehat{\boldsymbol{\beta}_p}^{(k)},\mathbf{b}_i^{(\ell,k)})-\vartheta_p\mathscr{E}(\mathbf{u}_i)^{(\ell,k)}\right]\right]\right\}\right],\end{aligned}$$

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

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 draw samples $\mathbf{b}_i^{(\ell,k)}$, $\ell=1,\ldots,m$, from $f(\mathbf{b}_i|\boldsymbol{\theta}^{(k)},\mathbf{y}_i)$, where m is the number of Monte Carlo simulations to be used, a number suggested to be large enough. A simulation method to draw samples from $f(\mathbf{b}_i|\boldsymbol{\theta}^{(k)},\mathbf{y}_i)$, is described in Subsection 4.2.

3.2 A SAEM algorithm

As mentioned in Subsection 2.2, the SAEM circumvents the cumbersome problem of simulating a large number of missing values at every iteration, leading to a faster and efficient solution than the MCEM. 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 for the conditional expectations by their stochastic approximations, given by

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

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

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_p^{(k+1)} &= \widehat{\boldsymbol{\beta}}_p^{(k)} + \left[\sum_{i=1}^{n} S_{1,i}^{(k)}\right]^{-1} \sum_{i=1}^{n} S_{2,i}^{(k)}, \\
\widehat{\sigma}^{(k+1)} &= \frac{1}{3N\tau_p^2} \sum_{i=1}^{n} S_{3,i}^{(k)}, \\
\widehat{\Psi}^{(k+1)} &= \frac{1}{n} \sum_{i=1}^{n} S_{4,i}^{(k)}.
\end{aligned} \tag{17}$$

Given a set of suitable initial values $\widehat{\boldsymbol{\theta}}^{(0)}$ (as detailed Appendix A), the SAEM iterates till convergence at iteration k if

$$\max_i \left\{ \frac{|\widehat{\theta}_i^{(k+1)} - \widehat{\theta}_i^{(k)}|}{|\widehat{\theta}_i^{(k)}| + \delta_1} \right\} < \delta_2 \tag{18}$$

is satisfied for three consecutive times where δ_1 and δ_2 are some small values pre established. The consecutive evalution of (18) avoids a fake convergence produced by an unlucky Monte Carlo simulation. Based on (Searle *et al.*, 1992) pag. 269, we use $\delta_1 = 0.001$ and $\delta_2 = 0.0001$ as suggested by several researchers. The proposed criterion above will need an extreme large number of iterations (more than usual) in order to detect convergence for parameters that are close to the boundary of the parametric space. In this case for variance components, a parameter value close to zero will inflate the ratio in (18) and the convergence will not be attained even though the likelihood was maximized with few iterations. As proposed by (Booth & Hobert, 1999) we use also a second convergence criteria besides to the first one, defined by

$$\max_i \left\{ \frac{|\widehat{\theta}_i^{(k+1)} - \widehat{\theta}_i^{(k)}|}{\sqrt{\widehat{\mathrm{var}}(\theta_i^{(k)})} + \delta_1} \right\} < \delta_2, \tag{19}$$

where (19) evaluates the parameter estimates changes relative to their standard errors leading to a convergence detection even for bounded parameters. Also the values δ_1 and δ_2 are some small values pre established and not necessarily equal to the one for (18). Based on simulation we suggest to fix $\delta_1 = 0.0001$ and to test different values for δ_2 between 0.0001 and 0.0005 when smaller means more accuracy. We use $\delta_1 = 0.0001$ and $\delta_2 = 0.0002$ by default which assures us a high accuracy. This stopping criteria is similar to the one proposed by (Bates & Watts, 1981) for Non linear Least Squares.

3.3 Missing data simulation method

In order to draw samples from $f(\mathbf{b}_i|\mathbf{y}_i, \boldsymbol{\theta})$, we utilize the Metropolis-Hastings (MH) algorithm (Metropolis *et al.*, 1953; Hastings, 1970), a MCMC algorithm for obtaining a sequence of random samples from a probability distribution for which direct sampling is not possible. The MH algorithm proceeds as follows:

Given $\boldsymbol{\theta} = \boldsymbol{\theta}^{(k)}$, for $i = 1, \ldots, n$;

1. Start with an initial value $\mathbf{b}_i^{(0,k)}$.

2. Draw $\mathbf{b}_i^* \sim h(\mathbf{b}_i^*|\mathbf{b}_i^{(\ell-1,k)})$ from a proposal distribution with the same support as the objective distribution $f(\mathbf{b}_i|\boldsymbol{\theta}^{(k)}, \mathbf{y}_i)$.

3. Generate $U \sim U(0,1)$.

4. If $U > \min\left\{1\,,\, \frac{f\left(\mathbf{b}_i^*|\boldsymbol{\theta}^{(k)},\mathbf{y}_i\right)h\left(\mathbf{b}_i^{(0,k)}|\mathbf{b}_i^*\right)}{f\left(\mathbf{b}_i^{(0,k)}|\boldsymbol{\theta}^{(k)},\mathbf{y}_i\right)h\left(\mathbf{b}_i^*|\mathbf{b}_i^{(0,k)}\right)}\right\}$, return to the step 2, else $\mathbf{b}_i^{(\ell,k)} = \mathbf{b}_i^*$

5. Repeat steps 2-4 until m samples $(\mathbf{b}_i^{(1,k)}, \mathbf{b}_i^{(2,k)}, \ldots, \mathbf{b}_i^{(m,k)})$ are drawn from $\mathbf{b}_i|\boldsymbol{\theta}^{(k)}, \mathbf{y}_i$.

Note that the marginal distribution $f(\mathbf{b}_i|\mathbf{y}_i, \boldsymbol{\theta})$ (omitting $\boldsymbol{\theta}$) can be represented as

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

where $\mathbf{b}_i \sim N_q(\mathbf{0}, \boldsymbol{\Psi})$ and $f(\mathbf{y}_i|\mathbf{b}_i) = \prod_{j=1}^{n_i} f(y_{ij}|\mathbf{b}_i)$, with $y_{ij}|\mathbf{b}_i \sim ALD\left(\mathbf{x}_{ij}^\top \boldsymbol{\beta}_p + \mathbf{z}_{ij}\mathbf{b}_i, \sigma, p\right)$. 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 the mean and variance-covariance matrix that are the stochastic approximations of the conditional expectation $\mathrm{E}(\mathbf{b}_i^{(k-1)}|\mathbf{y}_i)$ and the conditional variance $\mathrm{Var}(\mathbf{b}_i^{(k-1)}|\mathbf{y}_i)$ respectively, obtained from the last iteration of the SAEM algorithm. This candidate (with possible information about the shape of the target distribution) leads to better acceptance rate, and consequently a faster algorithm. The resulting chain $\mathbf{b}_i^{(1,k)}, \mathbf{b}_i^{(2,k)}, \ldots, \mathbf{b}_i^{(m,k)}$ is a MCMC sample from the marginal conditional distribution $f(\mathbf{b}_i|\boldsymbol{\theta}^{(k)}, \mathbf{y}_i)$. Due the dependent nature of these MCMC samples, at least 10 MC simulations are suggested.

4 Estimation

4.1 Likelihood Estimation

Given the observed data, the likelihood function $\ell_o(\boldsymbol{\theta}|\mathbf{y})$ of the model defined in (9)-(10) is given by

$$\ell_o(\boldsymbol{\theta}|\mathbf{y}) = \sum_{i=1}^{n} \log f(\mathbf{y}_i|\boldsymbol{\theta})) = \sum_{i=1}^{n} \log \int_{\mathbb{R}^q} f(\mathbf{y}_i|\mathbf{b}_i;\boldsymbol{\theta})\, f(\mathbf{b}_i;\boldsymbol{\theta})\, d\mathbf{b}_i, \tag{20}$$

where the integral can be expressed as an expectation with respect to $\mathbf{b}_i$, i.e., $E_{\mathbf{b}_i}[f(\mathbf{y}_i|\mathbf{b}_i;\theta)]$. 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 (22) as

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

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

$$\ell_o(\boldsymbol{\theta}|\mathbf{y}) \approx \sum_{i=1}^{n} \log \left\{ \frac{1}{m} \sum_{\ell=1}^{m} \left[\prod_{j=1}^{n_i} [f(y_{ij}|\mathbf{b}_i^{*(\ell)};\boldsymbol{\theta})] \frac{f(\mathbf{b}_i^{*(\ell)};\boldsymbol{\theta})}{\widehat{f}(\mathbf{b}_i^{*(\ell)};\boldsymbol{\theta})} \right] \right\}, \tag{21}$$

where $\{\mathbf{b}_i^{*(\ell)}\}$, $l = 1,\ldots,m$, is a MC sample from $\widehat{f}(\mathbf{b}_i^*;\boldsymbol{\theta})$, and $f(\mathbf{y}_i|\mathbf{b}_i^{*(\ell)};\boldsymbol{\theta})$ is expressed as $\prod_{j=1}^{n_i} f(y_{ij}|\mathbf{b}_i^{*(\ell)};\boldsymbol{\theta})$ due to independence. An efficient choice for $\widehat{f}(\mathbf{b}_i^{*(\ell)};\theta)$ is $f(\mathbf{b}_i|\mathbf{y}_i)$. Therefore, we use the same proposal distribution discussed in Subsection 4.2, and generate samples $\mathbf{b}_i^{*(\ell)} \sim N_q(\widehat{\boldsymbol{\mu}}_{\mathbf{b}_i}, \widehat{\boldsymbol{\Sigma}}_{\mathbf{b}_i})$, where $\widehat{\boldsymbol{\mu}}_{\mathbf{b}_i} = \mathrm{E}(\mathbf{b}_i^{(w)}|\mathbf{y}_i)$ and $\widehat{\boldsymbol{\Sigma}}_{\mathbf{b}_i} = \mathrm{Var}(\mathbf{b}_i^{(w)}|\mathbf{y}_i)$, which are estimated empirically during the last few iterations of the SAEM at convergence.

4.2 Missing data simulation method

In order to draw samples from $f(\mathbf{b}_i|\mathbf{y}_i,\boldsymbol{\theta})$, we utilize the Metropolis-Hastings (MH) algorithm (Metropolis *et al.*, 1953; Hastings, 1970), a MCMC algorithm for obtaining a sequence of random samples from a probability distribution for which direct sampling is not possible. The MH algorithm proceeds as follows:

Given $\boldsymbol{\theta} = \boldsymbol{\theta}^{(k)}$, for $i = 1,\ldots,n$;

1. Start with an initial value $\mathbf{b}_i^{(0,k)}$.
2. Draw $\mathbf{b}_i^* \sim h(\mathbf{b}_i^*|\mathbf{b}_i^{(\ell-1,k)})$ from a proposal distribution with the same support as the objective distribution $f(\mathbf{b}_i|\boldsymbol{\theta}^{(k)},\mathbf{y}_i)$.
3. Generate $U \sim U(0,1)$.
4. If $U > \min\left\{1, \frac{f\left(\mathbf{b}_i^*|\boldsymbol{\theta}^{(k)},\mathbf{y}_i\right)h\left(\mathbf{b}_i^{(0,k)}|\mathbf{b}_i^*\right)}{f\left(\mathbf{b}_i^{(0,k)}|\boldsymbol{\theta}^{(k)},\mathbf{y}_i\right)h\left(\mathbf{b}_i^*|\mathbf{b}_i^{(0,k)}\right)}\right\}$, return to the step 2, else $\mathbf{b}_i^{(\ell,k)} = \mathbf{b}_i^*$
5. Repeat steps 2-4 until m samples $(\mathbf{b}_i^{(1,k)}, \mathbf{b}_i^{(2,k)}, \ldots, \mathbf{b}_i^{(m,k)})$ are drawn from $\mathbf{b}_i|\boldsymbol{\theta}^{(k)},\mathbf{y}_i$.

Note that the marginal distribution $f(\mathbf{b}_i|\mathbf{y}_i,\boldsymbol{\theta})$ (omitting $\boldsymbol{\theta}$) can be represented as

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

where $\mathbf{b}_i \sim N_q(\mathbf{0},\boldsymbol{\Psi})$ and $f(\mathbf{y}_i|\mathbf{b}_i) = \prod_{j=1}^{n_i} f(y_{ij}|\mathbf{b}_i)$, with $y_{ij}|\mathbf{b}_i \sim ALD\big(\eta(\mathbf{A}_i\boldsymbol{\beta}_p + \mathbf{B}_i\mathbf{b}_i,\mathbf{x}_{ij}),\sigma,p\big)$. 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 the mean and variance-covariance matrix that are the stochastic approximations of the conditional expectation $\mathrm{E}(\mathbf{b}_i^{(k-1)}|\mathbf{y}_i)$ and the conditional variance $\mathrm{Var}(\mathbf{b}_i^{(k-1)}|\mathbf{y}_i)$ respectively, obtained from the last iteration of the SAEM algorithm. This candidate (with possible information about the shape of the target distribution) leads to better acceptance rate, and consequently a faster algorithm. The resulting chain $\mathbf{b}_i^{(1,k)}, \mathbf{b}_i^{(2,k)}, \ldots, \mathbf{b}_i^{(m,k)}$ is a MCMC sample from the marginal conditional distribution $f(\mathbf{b}_i|\theta^{(k)},\mathbf{y}_i)$. Due the dependent nature of these MCMC samples, at least 10 MC simulations are suggested.

5 Estimation

5.1 Likelihood Estimation

Given the observed data, the likelihood function $\ell_o(\boldsymbol{\theta}|\mathbf{y})$ of the model defined in (9)-(10) is given by

$$\ell_o(\boldsymbol{\theta}|\mathbf{y}) = \sum_{i=1}^{n} \log f(\mathbf{y}_i|\boldsymbol{\theta})) = \sum_{i=1}^{n} \log \int_{\mathbb{R}^q} f(\mathbf{y}_i|\mathbf{b}_i;\boldsymbol{\theta})\, f(\mathbf{b}_i;\boldsymbol{\theta})\, d\mathbf{b}_i, \tag{22}$$

where the integral can be expressed as an expectation with respect to $\mathbf{b}_i$, i.e., $E_{\mathbf{b}_i}[f(\mathbf{y}_i|\mathbf{b}_i;\theta)]$. 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 (22) as

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

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

$$\ell_o(\boldsymbol{\theta}|\mathbf{y}) \approx \sum_{i=1}^{n} \log \left\{ \frac{1}{m} \sum_{\ell=1}^{m} \left[\prod_{j=1}^{n_i} [f(y_{ij}|\mathbf{b}_i^{*(\ell)};\boldsymbol{\theta})]\, \frac{f(\mathbf{b}_i^{*(\ell)};\boldsymbol{\theta})}{\widehat{f}(\mathbf{b}_i^{*(\ell)};\boldsymbol{\theta})} \right] \right\}, \tag{23}$$

where $\{\mathbf{b}_i^{*(\ell)}\}$, $l = 1,\ldots,m$, is a MC sample from $\widehat{f}(\mathbf{b}_i^*;\boldsymbol{\theta})$, and $f(\mathbf{y}_i|\mathbf{b}_i^{*(\ell)};\boldsymbol{\theta})$ is expressed as $\prod_{j=1}^{n_i} f(y_{ij}|\mathbf{b}_i^{*(\ell)};\boldsymbol{\theta})$ due to independence. An efficient choice for $\widehat{f}(\mathbf{b}_i^{*(\ell)};\theta)$ is $f(\mathbf{b}_i|\mathbf{y}_i)$. Therefore, we use the same proposal distribution discussed in Subsection 4.2, and generate samples $\mathbf{b}_i^{*(\ell)} \sim N_q(\widehat{\boldsymbol{\mu}}_{\mathbf{b}_i}, \widehat{\boldsymbol{\Sigma}}_{\mathbf{b}_i})$, where $\widehat{\boldsymbol{\mu}}_{\mathbf{b}_i} = \mathrm{E}(\mathbf{b}_i^{(w)}|\mathbf{y}_i)$ and $\widehat{\boldsymbol{\Sigma}}_{\mathbf{b}_i} = \mathrm{Var}(\mathbf{b}_i|\mathbf{y}_i)$, which are estimated empirically during the last few iterations of the SAEM at convergence.

5.2 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

$$\mathbf{I}_e(\boldsymbol{\theta}|\mathbf{y}) = \sum_{i=1}^{n} \mathbf{s}(\mathbf{y}_i|\boldsymbol{\theta})\, \mathbf{s}^\top(\mathbf{y}_i|\widehat{\boldsymbol{\theta}}) - \frac{1}{n}\mathbf{S}(\mathbf{y}|\boldsymbol{\theta})\, \mathbf{S}^\top(\mathbf{y}|\boldsymbol{\theta}), \tag{24}$$

where $\mathbf{S}(\mathbf{y}|\boldsymbol{\theta}) = \sum_{i=1}^{n} \mathbf{s}(\mathbf{y}_i|\boldsymbol{\theta})$ and $\mathbf{s}(\mathbf{y}_i|\boldsymbol{\theta})$ is the empirical score function for the i-th individual. Replacing $\boldsymbol{\theta}$ by its ML estimator $\hat{\boldsymbol{\theta}}$ and considering $\boldsymbol{\nabla}_o(\hat{\boldsymbol{\theta}}) = \mathbf{0}$, equation (24) takes the simple form

$$\mathbf{I}_e(\widehat{\boldsymbol{\theta}}|\mathbf{y}) = \sum_{i=1}^{n} \mathbf{s}(\mathbf{y}_i|\widehat{\boldsymbol{\theta}})\, \mathbf{s}^\top(\mathbf{y}_i|\widehat{\boldsymbol{\theta}}). \tag{25}$$

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

$$\mathrm{s}(\mathbf{y}_i|\boldsymbol{\theta})^{(k)} = \mathbf{s}(\mathbf{y}_i|\boldsymbol{\theta})^{(k-1)} + \delta_k \left[\frac{1}{m} \sum_{\ell=1}^{m} \mathbf{s}(\mathbf{y}_i, \mathbf{q}^{(k,\ell)}; \boldsymbol{\theta}^{(k)}) - \mathbf{s}(\mathbf{y}_i|\boldsymbol{\theta})^{(k-1)} \right], \tag{26}$$

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 A.

6 Simulated data

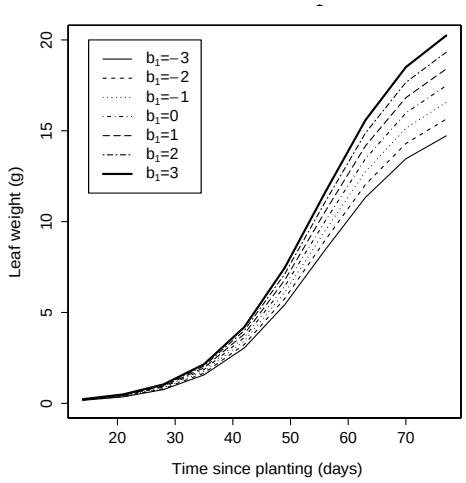
In order to examine the performance of the proposed method, here we present some simulation studies. The first simulation study shows that the ML estimates based on the SAEM algorithm do provide good asymptotic properties. The second study investigates the consequences for population inferences when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

6.1 Asymptotic properties

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

$$y_{ij} = \frac{\beta_1 + b_{1i}}{1 + \exp\left(-[t_{ij} - \beta_2]/\beta_3\right)} + \varepsilon_{ij},\ i = 1,\ldots,n,\ \ j = 1,\ldots,10, \tag{27}$$

Figure 2. Illustration of the effect of including the random effect $b1_i$ in the first parameter of the nonlinear growth-curve logistic model.

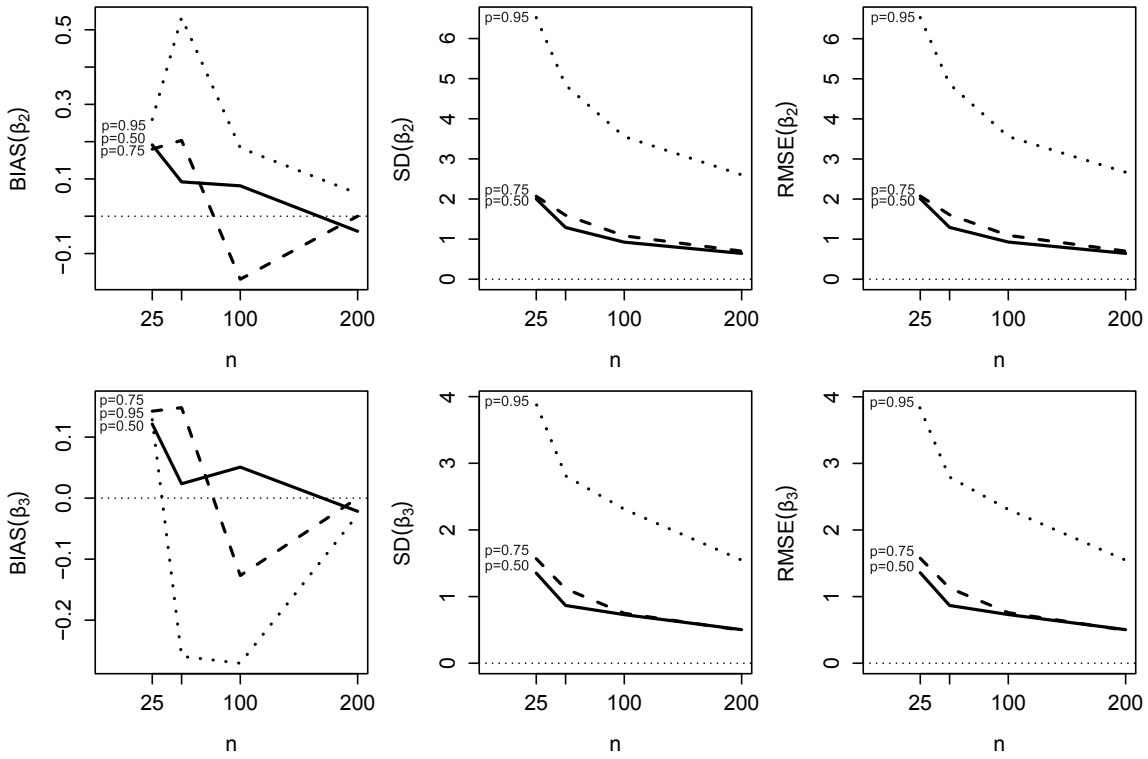


where $t_{ij} = 100, 267, 433, 600, 767, 933, 1100, 1267, 1433, 1600$ for all i. The goal is to estimate the fixed effects parameters β's for a grid of percentiles $p = \{0.50, 0.75, 0.95\}$. A random effects b_{1i} was added to the first growth parameter β_1 and its effect over the growth-curve is shown in Figure 4. Parameters interpretation for this model is going to be discussed in the Application Section. The random effects b_{1i} and the error $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1} \ldots, \varepsilon_{i10})^\top$ are non-correlated been $b_{1i} \overset{\text{iid}}{\sim} N(0, \sigma_b^2)$ and $\varepsilon_{ij} \overset{\text{iid}}{\sim} ALD(0, \sigma_e, p)$. We set $\boldsymbol{\beta}_p = (\beta_1, \beta_2, \beta_3)^\top = (200, 700, 350)^\top$, $\sigma_e = 0.5$, $\sigma_b^2 = 10$. Using the notation in (8) the matrices $\mathbf{A}_i$ and $\mathbf{B}_i$ are given by $\mathbf{I}_3$ and $(1,0,0)^\top$ respectively. For varying sample sizes of $n = 25, 50, 100$ and 200, we generate 100 data samples for each scenario. In addition, we also choose $m = 20$, $c = 0.25$ and $W = 500$ for the SAEM convergence parameters. For all scenarios, we compute the square root of the mean square error (RMSE), the bias (Bias) and the Monte carlo standard deviation (MC-Sd) for each parameter over the 100 replicates. They are defined as

$$\text{MC-Sd}(\widehat{\theta}_i) = \sqrt{\frac{1}{99}\sum_{j=1}^{100}\left(\widehat{\theta}_i^{(j)} - \overline{\widehat{\theta}_i}\right)^2} \qquad \text{and} \qquad \text{Bias}(\widehat{\theta}_i) = \overline{\widehat{\theta}_i} - \theta_i \tag{28}$$

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, for the bias we can see a patterns of convergence to zero when n increases for both parameters.

Figure 3. Bias, Standard Deviation and RMSE for β_1 (upper panel) and β_2 (lower panel) for varying 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 where it is evident that for extreme quantiles estimating, the standard deviation is much higher while for quantiles $q = 50$ and $q = 75$ are asymptotically equal. The worst scenario seems to happen while estimating extreme quantiles and maybe a sample size greater than 200 is needed to obtain a reasonably reduction of bias and SD. However, as a general rule, we can say that bias and MSE tend to approach to zero when the sample size is increasing, indicating that the approximates ML estimates based on the proposed SAEM algorithm do provide good asymptotic properties. The parameter β_1 has been discarded in the graphical analysis because it varies along quantiles so its bias too as seen in Table 1. This parameter represents the asymptotic growth so this parameter is highly susceptible to the quantile to be estimated, however it also provides good asymptotic properties for its standard deviation. Table 1 also show an excellent recovery for the nuisance parameter σ_e, small standard deviations and good asymptotic properties in terms of bias and SD.

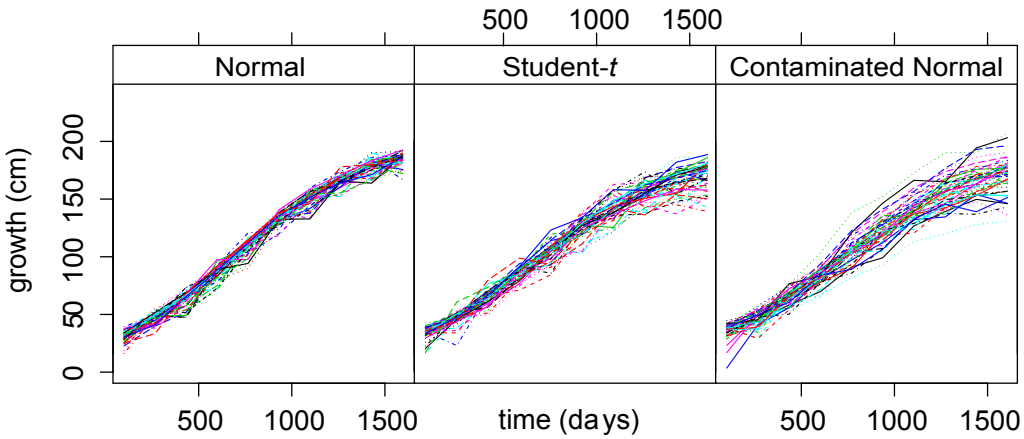
Table 1. Results based on 100 simulated samples. Monte carlo mean and standard deviation (MC Mean and MC-Sd) for the fixed effects β_1. β_2. β_3 and the nuisance parameter σ_e. obtained after fitting the QR-NLMM model under different settings of quantiles and sample sizes.

		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)

6.2 Robustness study

The goal of this simulation study is to asses the robustness or bias incurred when one assumes a normal distribution for random effects and the actual distribution belongs to a heavy tailed distributions. The use of heavy tailed distributions for the random effects will let us to simulate the presence of outliers leading us to test adequately the performance of the proposed method in terms of robustness. The design of this simulation study is as in the previous subsection but for a set of quantiles $\{0.50, 0.75\}$ and a fixed sample size $n = 50$ we are going to simulate 100 Monte Carlo samples generating the random effect term from a Student-t distribution with $\nu = 4$ degrees of freedom and from a Normal Contaminated distribution ($\nu_1 = 0.1, \nu_2 = \{0.1, 0.2, 0, 3\}$), i.e., with three scenarios of contamination, 10%, 20% and 30%. All simulations are created by using the same values of $\boldsymbol{\beta}_p = (200, 700, 350)^\top$, nuisance parameter $\sigma_e = 0.5$ and scale parameter $\sigma_b^2 = 10$ for the respectively random effect distribution.

Figure 4. Illustration of 50 simulated curves from the growth-curve logistic model using different distributions for the random effect term. From left to right panel, the random effects has been generated from a Normal, a Student t_4 and a Contaminated Normal($\nu_1 = 0.1, \nu_2 = 0.1$), all with location parameter $\mu = 0$ and scale parameter $\sigma_b^2 = 10$.



From Table 2 we can see that the proposed model is really robust even for worst scenarios of contamination. The parameter recovery is highly accurate even for the non-centered quantile 0.75. For quantile 0.75, the β_1 parameter tends to increase for higher levels of contamination. As expected,

Table 2. Results based on 100 simulated samples. MC Mean, Bias, MC-Sd and RMSE for the fixed effects β_1, β_2, β_3 and the nuisance parameter σ_e obtained after fitting the QR-NLMM for quantiles 0.50 and 0.75 using four different distribution settings for the random effects.

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

the MC-Sd and consequently the RMSE increase in presence of outliers. As a general rule, we can conclude that the proposed model is robust in presence of outliers or misspecification of the random effect distribution.

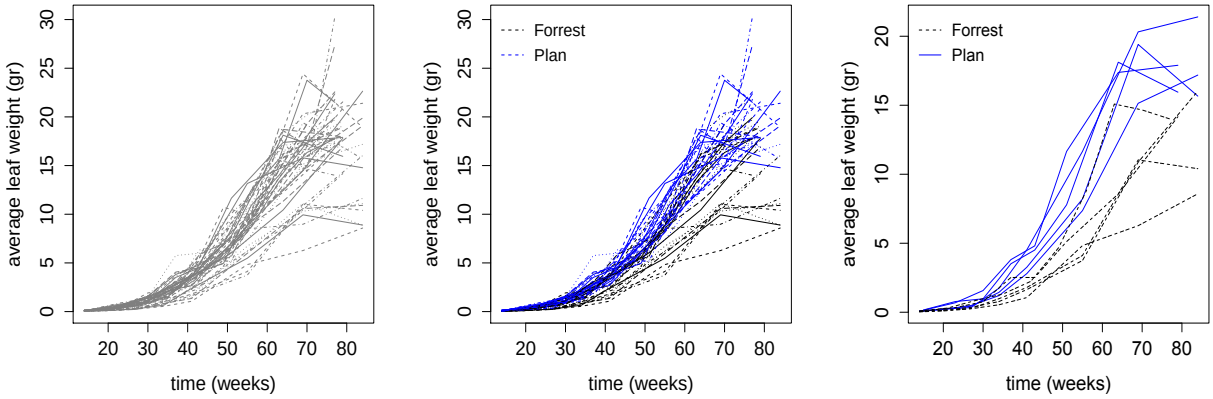
7 Illustrative examples

In this section, we illustrate the application of our method to two interesting longitudinal datasets from the literature.

7.1 Growth curve: Soybean data

For the first application, we are going to consider the Soybean genotypes data analyzed by Davidian & Giltinan (1995) and Pinheiro & Bates (2000), a longitudinal experiment consisting of measuring along time the leaf weight (in g) as a measure of growth of two kinds of Soybean genotype plants to be compared, a commercial variety, Forrest (F), and an experimental strain, Plan Introduction #416937 (P). The samples were taken approximately weekly during 8 to 10 weeks. For three consecutive years, 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.

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.



We use the three parameter logistic model in (27) introducing a random effect term for each parameter and a dichotomic covariate as

$$y_{ij} = \frac{\varphi_{1i}}{1+\exp\left(-[t_{ij}-\varphi_{2i}]/\varphi_{3i}\right)} + \varepsilon_{ij},\ i = 1,\ldots,412,\ \ j = 1,\ldots,n_i, \tag{29}$$

where,

$$\begin{aligned}
\varphi_{1i} &= \beta_1 + \beta_4 gen_i + b_{1i} \\
\varphi_{2i} &= \beta_2 + b_{2i} \\
\varphi_{3i} &= \beta_3 + b_{3i}.
\end{aligned}$$

The observed value y_{ij} represents mean weight of leaves (in g) from six randomly selected soybean plants in the ith plot, after t_{ij} days of been planted; gen_i is a dichotomic variable for the genotype of plant i (0=forrest, 1=plan Introduction) and ε_{ij} is the measurement error for the 412 plants. Let be $\boldsymbol{\beta}_p = (\beta_1, \beta_2, \beta_3, \beta_4)^\top$ and $\mathbf{b}_i = (b_{1i}, b_{2i}, b_{3i})^\top$ the fixed and random effects vector respectively. Then the matrices $\mathbf{A}_i$ and $\mathbf{B}_i$ are defined as

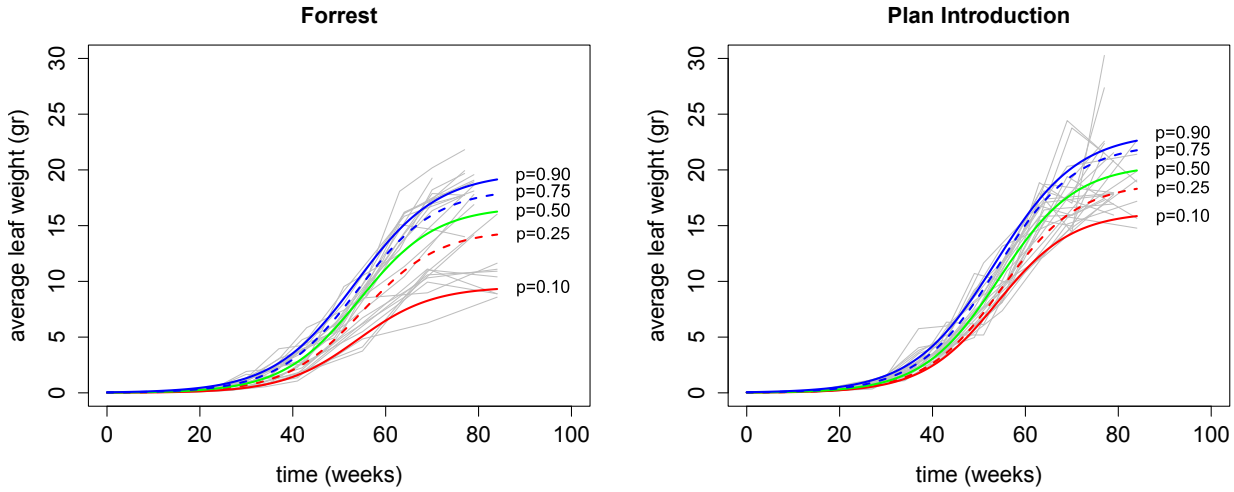
$$\mathbf{A}_i = \begin{pmatrix} 1 & 0 & 0 & gen_i \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \qquad \text{and} \qquad \mathbf{B}_i = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{30}$$

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. Due the goal of comparing the final (asymptotic) growth of the two kind of Soybeans, the dichotomic covariate gen_i was incorporated in the first component of the growth function, then the fourth fixed effect β_4 will represent the difference (in g) of the asymptotic leaf weight between the plan introduction type and the forrest one (control). As seen in middle and right panel of figure 5, it appears to exist a significance difference between the experimental and control Soybean so we expect a positive non zero β_4 estimate for most of quantiles.

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 clear how the extreme quantiles estimation functions captures the full data variability and evidences some atypical observations, specially for the plan introduction genotype. Quantile functions (for same quantile value) looks really different for each genotype due the significance of β_4 over the model as seen in Figure 7.

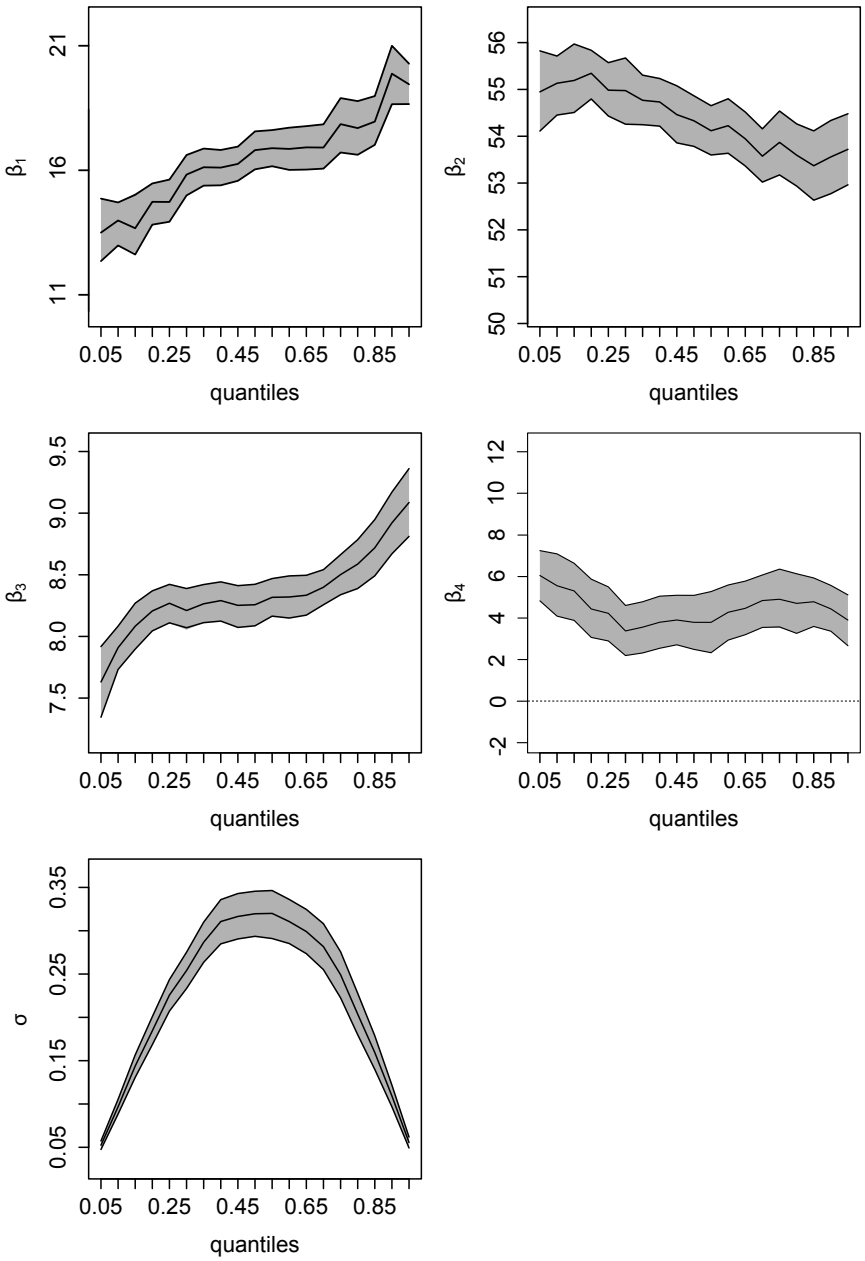
After fitting the quantile regression over the grid $p = \{0.05, 0.10, ..., 0.95\}$, we show a graphical summary of the obtained results in Figure 7. We assessed the convergence of the fixed effect estimates, variance components of the random effects and nuisance parameters using graphical criteria as shown in Figure 11 in Appendix D. It shows a 95% confidence band for the fixed effect

Figure 6. Fitted quantile regression for several quantiles for the Soybean data by genotype.



parameters β_1, β_2, β_3, β_4 and for the nuisance parameter σ where the solid lines are the $Q_{0.025}$ percentile and $Q_{0.975}$ percentile obtained through the estimation of the standard errors based on the empirical information matrix. We can see that the effect of the genotype results significant for all the quantile profile and the difference varies with respect to the conditional quantile been more significant for lower quantiles. This can be corroborated in Figure 6 where the difference between the 0.10 estimated quantile functions for different genotypes is greater than for other quantiles.

Figure 7. Point estimates (center solid line) and 95% confidence intervals for model parameters after fitting the QR to the Soybean data across various quantiles. The interpolated curves are spline-smoothed.



Using the information provided by the 95th percentile, we infer that the Soybean plants that grew more have a mean leaf weight around 19.35 grams for the Forrest genotype and 23.25 grams for the plan introduction one, then the asymptotic difference for the two genotypes is around 4 grams. The behavior of the estimate of the nuisance parameter σ is symmetric with respect to $p = 0.50$, taking its maximum value and variability on it and both decreasing for extreme quantiles. This behavior is because the variance within subjects depends of the quantile to be estimated, been proportional to the asymmetry of the error term then for extreme quantiles the nuisance parameter should be reduced.

7.2 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, we analyze the HIV viral load of 46 HIV-1 infected patients under antiretroviral treatment (protease inhibitor and reverse transcriptase inhibitor drugs). The viral load and some other covariates were mesured several times days after the start of treatment been 4 and 10 the minimum and maximum number of measures per patient respectively. Wu (2002) found that the only significance covariate for modelling the virus load was the CD4 therefore the other covariates even though they could be incorporated to the model for instance they are going to be discard. Figure 8 shows the profile of viral load in log10 scale and CD4 cell count/100 per cubic millimeter versus time (in days/100) for six randomly selected patients. We can see that appear to exist some relationship between the viral load and the CD4 cell count and it seems to be inversely proportional, 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 to invasion of viruses and/or bacteria so lower CD4 count means a weaker immune system. Normal counts of CD4 cells are from 500-1000 cells per cubic millimeter whereas fewer counts than 200 cells/mm^3 will be a high qualification to diagnose AIDS. We can evidence the mentioned before in the right panel of Figure 8 where the three patients who have less than 200 CD4 cells/mm^3 (delimited by the horizontal dashed line in 0.02) are the ones with higher levels of viral load.

In order to fit the nonlinear data we will use the nonlinear model proposed by Wu (2002) and also used by Lachos *et al.* (2013). The proposed bi-exponential NLME model is given by:

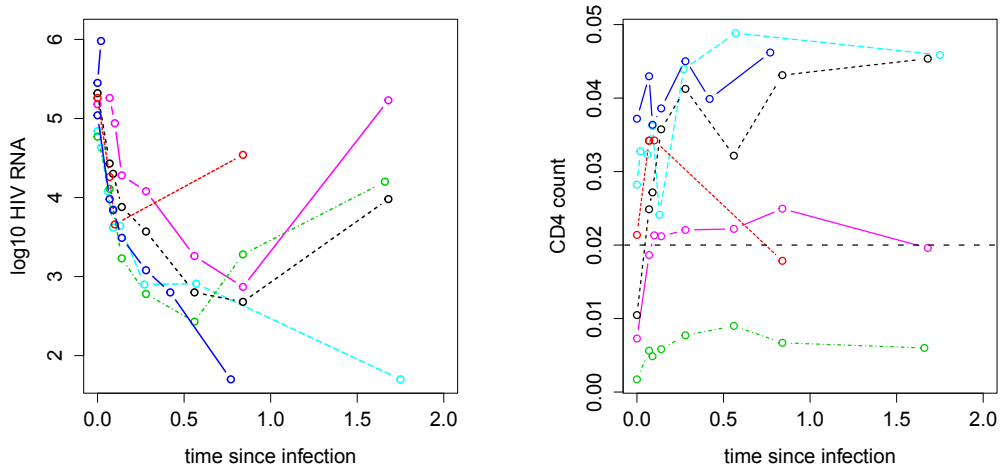
$$y_{ij} = \log_{10}\left(e^{(\varphi_{1i}-\varphi_{2i}t_{ij})} + e^{(\varphi_{3i}-\varphi_{4i}t_{ij})}\right) + \varepsilon_{ij}, \quad i = 1,\ldots,46,\ j = 1,\ldots,n_i, \tag{31}$$

with

$$\begin{aligned} \varphi_{1i} &= \beta_1 + b_{1i} & \varphi_{2i} &= \beta_2 + b_{2i} \\ \varphi_{3i} &= \beta_3 + b_{3i} & \varphi_{4ij} &= \beta_4 + \beta_5 CD4_{ij} + b_{4i}, \end{aligned}$$

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 for the 46 patients. Let be $\boldsymbol{\beta}_p = (\beta_1,\beta_2,\beta_3,\beta_4,\beta_5)^\top$ and $\mathbf{b}_i = (b_{1i},b_{2i},b_{3i},b_{4i})^\top$ the fixed and random effects vector respectively and $\mathbf{CD4}_i = (CD4_{i1},\ldots,CD4_{in_i})^\top$. Then the matrices $\mathbf{A}_i$ and $\mathbf{B}_i$ are defined as

Figure 8. ACTG 315 data. Profiles of viral load (response) in log10 scale and CD4 cell count (in cells/100mm^3) for ten randomly selected patients.



$$\mathbf{A}_i = \begin{pmatrix} \mathbf{I}_3 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_{n_i} & \mathbf{CD4}_i \end{pmatrix} \qquad \text{and} \qquad \mathbf{B}_i = \begin{pmatrix} \mathbf{I}_3 & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_{n_i} \end{pmatrix}. \tag{32}$$

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 (31) see Grossman *et al.* (1999) and Perelson *et al.* (1997).

Figure 9. ACTG 315 data: Fitted quantile regression functions overlayed for the HIV data.

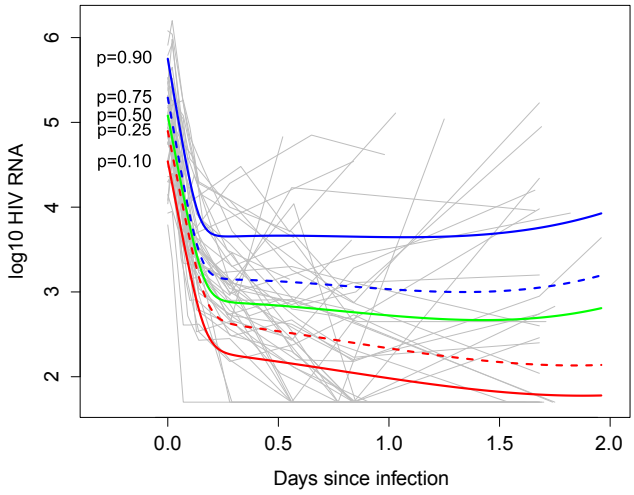
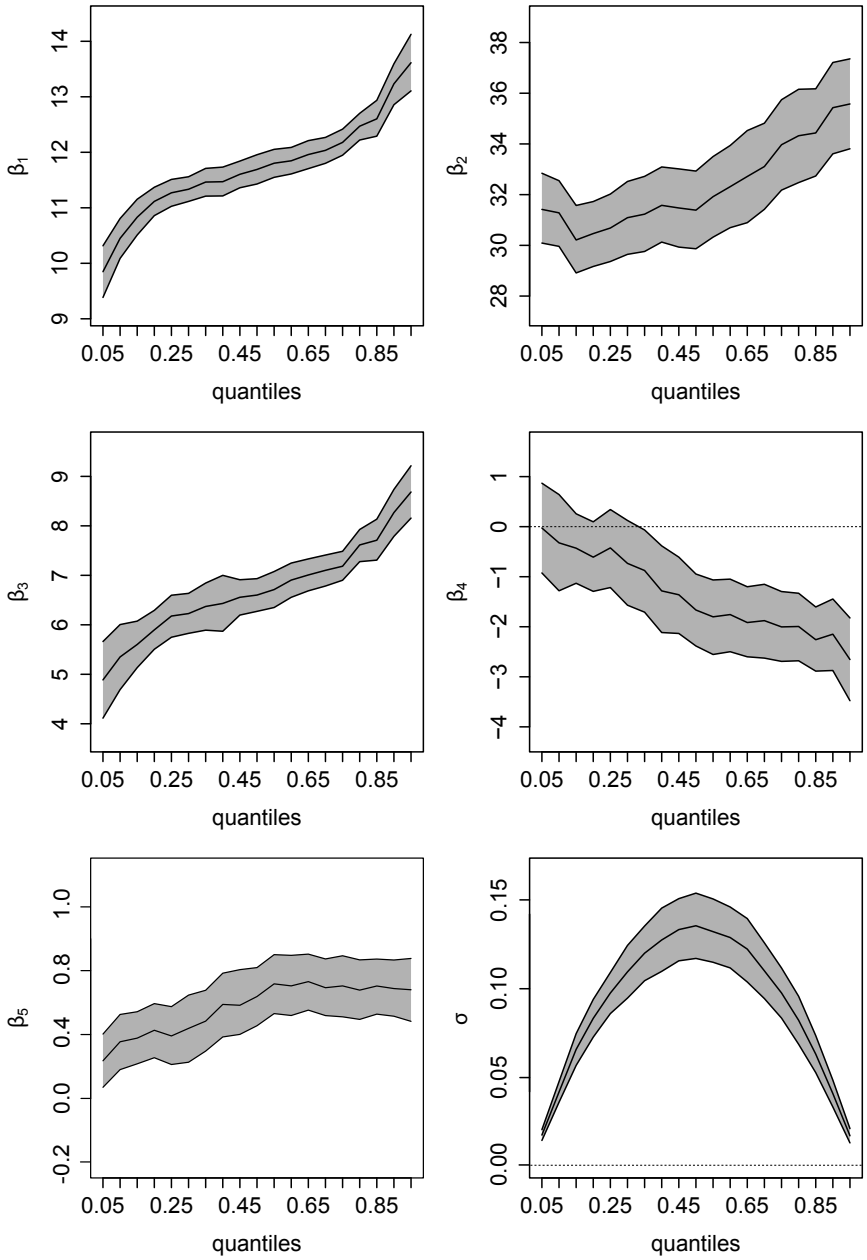


Figure 9 shows the fitted regression lines for quantiles 0.10, 0.25, 0.50, 0.75 and 0.90 for the HIV data. In order to plot, first, we fixed the CD4 covariate using the predicted sequence from a linear regression (including a quadratic term) for explaning the CD4 cell count with respect to time. We can see how quantile estimated functions follow the data behaviour satisfactorily 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.

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



The results after fitting QR over the grid of quantiles $p = \{0.05, 0.10, ..., 0.95\}$ are shown in figure 10. The convergence of estimates for all parameters were also assessed using the graphical criteria in Figure 12 in Appendix D. based on We have found that the first phase viral decay rate is positive and its effect tends to increase proportionally along quantiles. For the second phase viral decay rate we have that this second rate is positive correlated with the CD4 count and therefore with the therapy time. Then, 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 behavior 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 then its effect becomes constant for greater quantiles. The behavior of the estimate of the nuisance parameter σ is the same as in Application 1.

8 Conclusions

In this paper, we investigate quantile regression of nonlinear mixed effects models from a likelihood-based perspective. The ALD distribution and SAEM algorithm are combined to propose an exact ML estimation method, in contrast to the approximated method proposed by Geraci & Bottai (2014). 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 and applications to two real datasets. We believe that this paper is the first attempt for exact ML estimation in the context of QR-NLMMs. The methods developed can be readily implemented inside `R` through package `qrNLMM()`.

There are a number of possible extensions of the current work. For modelling both skewness and long tails in the random effects, the scale mixtures of skew-normal (SMSN) distributions (Lachos *et al.*, 2010) is a feasible choice. Also, HIV viral loads studies include covariates (viz. 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 305054/2011-2 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 $\mathbf{y}_i \sim ALD(\boldsymbol{\eta}(\boldsymbol{\beta}_p, \mathbf{0}), \sigma, p)$. 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-LMM 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 $u_{ik}|\, y_{ik}, \mathbf{b}_i$, for all $j,k = 1,2,\ldots,n_i$ and $j \neq k$, we can write $\mathbf{u}_i|\mathbf{y}_i, \mathbf{b}_i = [\ u_{i1}|y_{i1}, \mathbf{b}_i \quad u_{i2}|y_{i2}, \mathbf{b}_i \quad \cdots \quad u_{in_i}|y_{in_i}, \mathbf{b}_i\ ]^\top$. 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

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

We already have $u_{ij}|y_{ij}, \mathbf{b}_i \sim GIG(\frac{1}{2}, \chi_{ij}, \psi)$ where χ_{ij} and ψ are defined in (16). 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 (12), the complete log-likelihood function can be rewritten as

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

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

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

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

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

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

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

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 (25) by approximating the score for the observed log-likelihood by the stochastic approximation given in (26).

Appendix D Figures

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 for the Soybean data. The vertical dashed line delimits the beginning of the almost sure convergence as defined by the cut-point parameter $c = 0.25$.

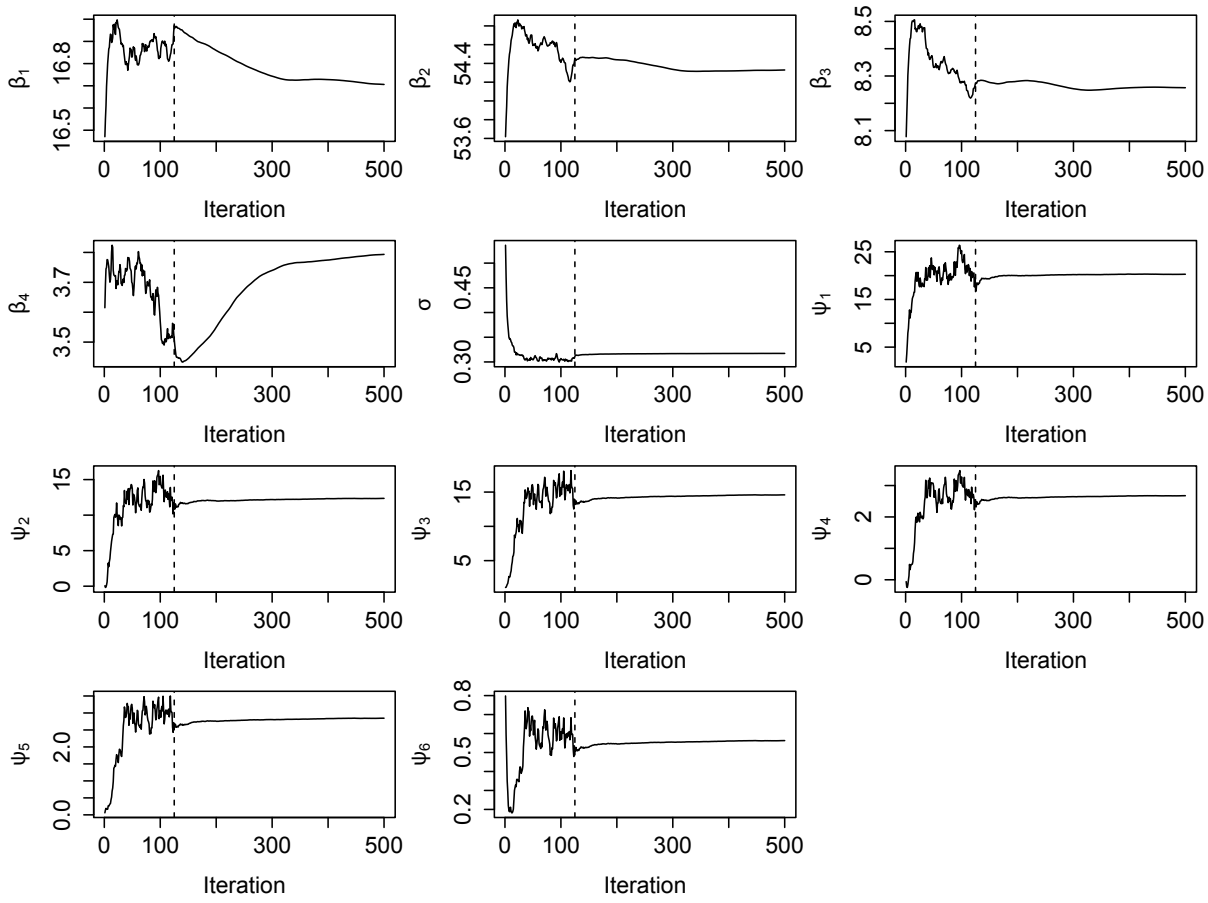
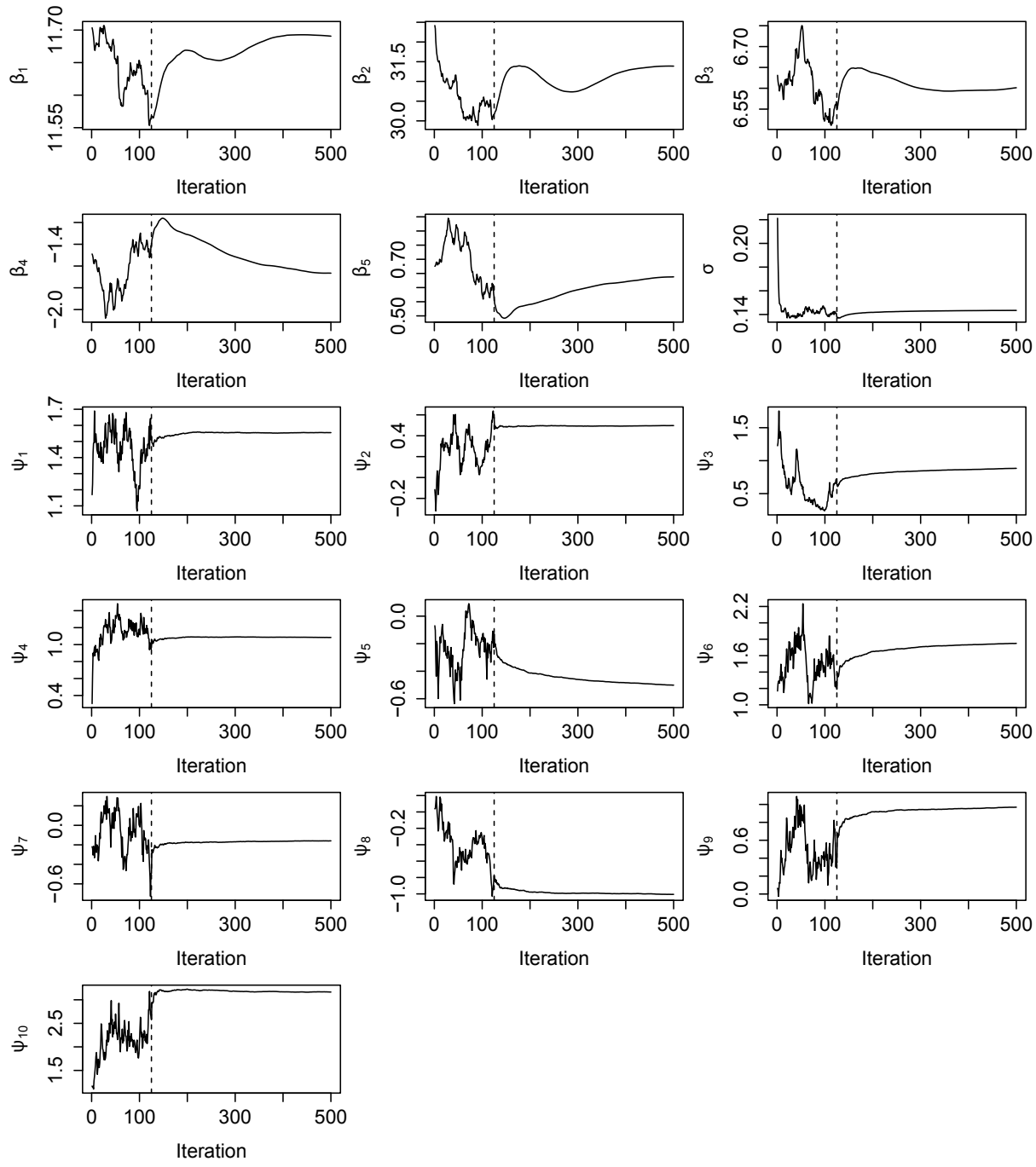


Figure 12. Graphical summary for the convergence of the fixed effect estimates, variance components of the random effects, and nuisance parameters performing a median regression for the HIV data. The vertical dashed line delimits the beginning of the almost sure convergence as defined by the cut-point 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, STÉPHANIE AND KUHN, ESTELLE AND TROUVÉ, ALAIN AND OTHERS (2010). Construction of Bayesian deformable models via a stochastic approximation algorithm: a Convergence study. *Bernoulli*, **16**(3), 641–678.

BARNDORFF-NIELSEN, OLE E AND SHEPHARD, NEIL (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.

BATES, DOUGLAS M AND WATTS, DONALD G (1981). A Relative Off set Orthogonality Convergence Criterion for Nonlinear least Squares. *Technometrics*, **23**(2), 179–183.

BOOTH, J. G. AND 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. AND GILTINAN, DAVID (2003). Nonlinear models for repeated measurement data: an overview and update. *Journal of Agricultural, Biological and Environmental Statistics*, **8**(4), 387–419.

DAVIDIAN, MARIE AND GILTINAN, DAVID M (1995). *Nonlinear Models for Repeated Measurement Data*, volume 62. CRC Press.

DELYON, BERNARD AND LAVIELLE, MARC AND MOULINES, ERIC (1999). Convergence of a stochastic approximation version of the EM algorithm. *Annals of Statistics*, **8**, 94–128.

A. DEMPSTER AND N. LAIRD AND D. RUBIN (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, **39**, 1–38.

FU, LIYA AND WANG, YOU-GAN (2012). Quantile regression for longitudinal data with a working correlation model. *Computational Statistics & Data Analysis*, **56**(8), 2526–2538.

GALVAO, ANTONIO F AND MONTES-ROJAS, GABRIEL V (2010). Penalized quantile regression for dynamic panel data. *Journal of Statistical Planning and Inference*, **140**(11), 3476–3497.

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

GERACI, MARCO AND BOTTAI, MATTEO (2007). Quantile regression for longitudinal data using the asymmetric Laplace distribution. *Biostatistics*, **8**(1), 140–154.

GERACI, MARCO AND BOTTAI, MATTEO (2014). Linear quantile mixed models. *Statistics and Computing*, **24**(3), 461–479.

GROSSMAN, ZVI AND POLIS, MICHAEL AND FEINBERG, MARK B AND GROSSMAN, ZEHAVA AND LEVI, ITSCHAK AND JANKELEVICH, SHIRLEY AND YARCHOAN, ROBERT AND BOON, JACOB AND DE WOLF, FRANK AND LANGE, JOEP MA AND OTHERS (1999). Ongoing hiv dissemination during haart. *Nature medicine*, **5**(10), 1099–1104.

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

HUANG, YANGXIN AND DAGNE, GETACHEW (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, ROGER (2004). Quantile regression for longitudinal data. *Journal of Multivariate Analysis*, **91**(1), 74–89.

KOENKER, ROGER (2005). *Quantile Regression*. Cambridge University Press, New York, NY.

KOTZ, S. AND KOZUBOWSKI, T.J. AND PODGORSKI, K. (2001). *The Laplace distribution and generalizations: A revisit with applications to communications, economics, engineering, and finance*. Birkhauser.

KUHN, ESTELLE AND LAVIELLE, MARC (2004). Coupling a stochastic approximation version of EM with an MCMC procedure. *ESAIM: Probability and Statistics*, **8**, 115–131.

KUHN, ESTELLE AND LAVIELLE, MARC (2005). Maximum likelihood estimation in nonlinear mixed effects models. *Computational Statistics & Data Analysis*, **49**(4), 1020–1038.

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

V. H. LACHOS AND P. GHOSH AND R. B. ARELLANO-VALLE (2010). Likelihood based Inference for Skew–Normal Independent Linear Mixed Models. *Statistica Sinica*, **20**(1), 303–322.

LACHOS, VICTOR H AND CASTRO, LUIS M AND DEY, DIPAK K (2013). Bayesian inference in nonlinear mixed-effects models using normal independent distributions. *Computational Statistics & Data Analysis*, **64**, 237–252.

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

LIPSITZ, STUART R AND FITZMAURICE, GARRETT M AND MOLENBERGHS, GEERT AND ZHAO, LUE PING (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, THOMAS 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, ISAAC (1989). A fast improvement to the EM algorithm on its own terms. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 127–138.

NICHOLAS METROPOLIS AND ARIANNA W. ROSENBLUTH AND MARSHALL N. ROSENBLUTH AND AUGUSTA H. TELLER AND EDWARD TELLER (1953). Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, **21**, 1087–1092.

MEZA, C. AND OSORIO, F. AND DE LA CRUZ, R. (2012). Estimation in nonlinear mixed-effects models using heavy-tailed distributions. *Statistics and Computing*, **22**, 121–139.

PERELSON, ALAN S AND ESSUNGER, PAULINA AND CAO, YUNZHEN AND VESANEN, MIKA AND HURLEY, ARLENE AND SAKSELA, KALLE AND MARKOWITZ, MARTIN AND HO, DAVID D (1997). Decay characteristics of hiv-1-infected compartments during combination therapy.

J.C. PINHEIRO AND D.M. BATES (1995). Approximations to the log-likelihood function in the nonlinear mixed effects model. *Journal of Computational and Graphical Statistics*, **4**, 12–35.

PINHEIRO, JOSÉ C AND BATES, DOUGLAS M (2000). *Mixed-effects Models in S and S-PLUS*. Springer, New York, NY.

SEARLE, SHAYLE R AND CASELLA, G AND MCCULLOCH, CE (1992). Variance components, 1992.

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

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

WEI, GREG CG AND TANNER, MARTIN 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.

WU, CF JEFF (1983). On the convergence properties of the em algorithm. *The Annals of statistics*, pages 95–103.

WU, LANG (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.

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

YU, K. AND MOYEED, R.A. (2001). Bayesian quantile regression. *Statistics & Probability Letters*, **54**(4), 437–447.

YU, KEMING AND ZHANG, JIN (2005). A three-parameter asymmetric Laplace distribution and its extension. *Communications in Statistics - Theory and Methods*, **34**(9-10), 1867–1879.

YUAN, YING AND YIN, GUOSHENG (2010). Bayesian quantile regression for longitudinal studies with nonignorable missing data. *Biometrics*, **66**(1), 105–114.