Quantile Regression for Linear Mixed Models: A Stochastic Approximation EM approach

Christian E. Galarza[a] Dipankar Bandyopadhyay[b]*
Victor H. Lachos[a]

[a]*Departamento de Estatística, Universidade Estadual de Campinas, Campinas, Brazil*
[b]*Division of Biostatistics, University of Minnesota, Minneapolis, MN, USA.*

August 13, 2015

Abstract

This paper develops a likelihood-based approach to analyze quantile regression (QR) models for continuous longitudinal data via the asymmetric Laplace distribution (ALD). Compared to the conventional mean regression approach, 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. Exploiting the nice hierarchical representation of the ALD, our classical approach follows a Stochastic Approximation of the EM (SAEM) algorithm in deriving exact maximum likelihood estimates of the fixed-effects and variance components. 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. Our empirical results clearly indicate that the SAEM estimates outperforms the estimates obtained via the combination of Gaussian quadrature and non-smooth optimization routines of the Geraci (2014)'s approach in terms of standard errors and mean square error. The proposed SAEM algorithm is implemented in the `R` package `qrLMM()`

Keywords: Asymmetric Laplace distribution; SAEM algorithm; `R` package `qrLMM`

1 Introduction

Linear mixed-effects models (LMM) are frequently used to analyze grouped/clustered data (such as longitudinal data, repeated measures, and multilevel data) because of their ability to handle within-subject correlations that characterizes grouped data (Pinheiro and Bates, 2000). Majority of these LMMs 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

*Address for correspondence: Division of Biostatistics, School of Public Health, University of Minnesota, A452 Mayo MMC 303, 420 Delaware Street SE, Minneapolis, MN 55455. E-mail: `dbandyop@umn.edu`.

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 distributional assumption on the error terms, except that the error term has a zero conditional quantile. Because of its popularity and the flexibility it provides, standard QR methods are implementable via available software packages, such as, the `R` package `quantreg`.

Although QR was initially developed under a univariate framework, the abundance of clustered data in recent times led to its extensions into mixed modeling framework (classical, or Bayesian) via either the distribution-free route (Lipsitz et al., 1997; Galvao and Montes-Rojas, 2010; Galvao Jr, 2011; Fu and Wang, 2012), or the traditional likelihood-based route mostly using the ALD (Geraci and Bottai, 2007; Yuan and Yin, 2010; Geraci and Bottai, 2014). Among the ALD-based models, Geraci and 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 and Bottai (2014) extended it 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.

Although the literature on QR-LMM is now substantial, there are no studies conducting exact inferences in the context of QR-LMM from a likelihood-based perspective. In this paper, we proceed to achieve that via a robust parametric ALD-based QR-LMM 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 and Bottai (2014). 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 and Lavielle (2005) showed that the SAEM algorithm is very efficient in computing the ML estimates in mixed effects models. Our empirical results using the SAEM are more efficient than the proposition of Geraci and Bottai (2014) for simulated data. Furthermore, application of our method to two longitudinal datasets is illustrated via the `R` package `qrLMM()`.

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 LMM, while Section 4 outlines the likelihood estimation and standard errors. Section 5 presents simulation studies to compare the finite sample performance of our proposed methods with the competing Geraci and Bottai (2014) method. Application of the SAEM method to two longitudinal datasets, one examining cholesterol level and the other on orthodontic distance growth 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 outline the EM and SAEM algorithms for ML estimation.

2.1 Connection between QR and ALD

Following Yu and 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 $W = \rho_p(\frac{Y-\mu}{\sigma})$ follows an exponential(1) distribution. Figure 1 plots the ALD illustrating how 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. The ALD abides by the following stochastic representation (Kotz et al., 2001; Kuzobowski and 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

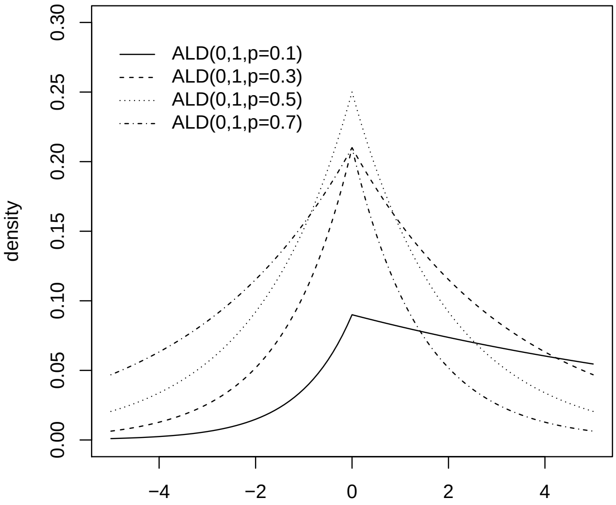

Figure 1: Standard asymmetric Laplace density

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

where $\vartheta_p = \frac{1-2p}{p(1-p)}$ and $\tau_p^2 = \frac{2}{p(1-p)}$, and $\overset{d}{=}$ 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 follows

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

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

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

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

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

The moments of U can be expressed as

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

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 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 $Q(\boldsymbol{\theta}|\widehat{\theta}^{(k)})=\mathrm{E}\Big\{\ell_c(\boldsymbol{\theta};\mathbf{y}_{\text{com}})|\widehat{\boldsymbol{\theta}}^{(k)},\mathbf{y}_{\text{obs}}\Big\}$, 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 $\boldsymbol{\theta}$ to obtain$\widehat{\boldsymbol{\theta}}^{(k+1)}$.

However, in some applications of the EM algorithm, the E-step cannot be obtained analytically and has to be calculated using simulations. Wei and 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 and 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 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 and 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 the algorithm with memory will converge slowly (almost sure convergence) to the ML solution. We suggested the following choice of the smoothing parameter:

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

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. 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 linear mixed models and algorithms

We consider the following general LMM $y_{ij} = \mathbf{x}_{ij}^\top \boldsymbol{\beta} + \mathbf{z}_{ij}\mathbf{b}_i + \varepsilon_{ij},\ i = 1,\ldots,n,\ j = 1,\ldots,n_i$, where y_{ij} is the jth measurement of a continuous random variable for the ith subject, $\mathbf{x}_{ij}^\top$ are row vectors of a known design matrix of dimension $N \times k$ corresponding to the $k \times 1$ vector of population-averaged fixed effects $\boldsymbol{\beta}$, $\mathbf{z}_{ij}$ is a $q \times 1$ design matrix corresponding to the $q \times 1$ vector of random effects $\mathbf{b}_i$, and ε_{ij} the independent and identically distributed random errors. We define pth quantile function of the response y_{ij} as

$$Q_p(y_{ij}|\mathbf{x}_{ij},\mathbf{b}_i) = \mathbf{x}_{ij}^\top \boldsymbol{\beta}_p + \mathbf{z}_{ij}\mathbf{b}_i. \tag{7}$$

where Q_p denotes the inverse of the unknown distribution function F, $\boldsymbol{\beta}_p$ is the regression coefficient corresponding to the pth quantile, 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 $\varepsilon_{ij} \sim ALD(0,\sigma)$. Then, $y_{ij}|\mathbf{b}_i$ independently follows as ALD with the density given by

$$f(y_{ij}|\boldsymbol{\beta}_p,\mathbf{b}_i,\sigma) = \frac{p(1-p)}{\sigma}\exp\left\{-\rho_p\left(\frac{y_{ij} - \mathbf{x}_{ij}^\top\boldsymbol{\beta}_p - \mathbf{z}_{ij}\mathbf{b}_i}{\sigma}\right)\right\}, \tag{8}$$

Using a MCEM algorithm, a QR-LMM with random intercepts $(q = 1)$ was proposed by Geraci and Bottai (2007). More recently, Geraci and Bottai (2014) extended that setup to accommodate multiple random effects where the estimation of fixed effects and covariance matrix of the random effects were accomplished via a combination of Gaussian quadrature approximations and non-smooth optimization algorithms. Here, we consider a more general correlated random effects framework with general dispersion matrix $\boldsymbol{\Psi} = \boldsymbol{\Psi}(\boldsymbol{\alpha})$.

3.1 A MCEM algorithm

First, we develop a MCEM algorithm for ML estimation of the parameters in the QR-LMM. From (3), the QR-LMM defined in (7)-(8) can be represented in a hierarchical form as:

$$\begin{aligned}
\mathbf{y}_i|\mathbf{b}_i,\mathbf{u}_i &\sim N_{n_i}\left(\mathbf{x}_i^\top\boldsymbol{\beta}_p + \mathbf{z}_i\mathbf{b}_i + \vartheta_p\mathbf{u}_i, \sigma\tau_p^2\mathbf{D}_i\right),\\
\mathbf{b}_i &\sim N_q\left(\mathbf{0},\boldsymbol{\Psi}\right),\\
\mathbf{u}_i &\sim \prod_{j=1}^{n_i}\exp(\sigma),
\end{aligned} \tag{9}$$

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 $\mathbf{y}_{ic} = (\mathbf{y}_i^\top,\mathbf{b}_i^\top,\mathbf{u}_i^\top)^\top$, 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 $\theta^{(k)} = (\boldsymbol{\beta}_p^{(k)\top},\sigma^{(k)},\boldsymbol{\alpha}^{(k)\top})^\top$, 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 $\ell_c(\boldsymbol{\theta};\mathbf{y}_c) = \sum_{i=1}^n \ell_c(\boldsymbol{\theta};\mathbf{y}_{ic})$, where

$$\begin{aligned}
\ell_c(\boldsymbol{\theta};\mathbf{y}_{ic}) &= \text{constant} - \frac{3}{2}n_i log\sigma - \frac{1}{2}log\left|\boldsymbol{\Psi}\right| - \frac{1}{2}\mathbf{b}_i^\top\boldsymbol{\Psi}^{-1}\mathbf{b}_i - \frac{1}{\sigma}\mathbf{u}_i^\top\mathbf{1}_{n_i}\\
&- \frac{1}{2\sigma\tau_p^2}(\mathbf{y}_i - \mathbf{x}_i^\top\boldsymbol{\beta}_p - \mathbf{z}_i\mathbf{b}_i - \vartheta_p\mathbf{u}_i)^\top\mathbf{D}_i^{-1}(\mathbf{y}_i - \mathbf{x}_i^\top\boldsymbol{\beta}_p - \mathbf{z}_i\mathbf{b}_i - \vartheta_p\mathbf{u}_i).
\end{aligned} \tag{10}$$

Given the current estimate $\boldsymbol{\theta} = \boldsymbol{\theta}^{(k)}$, the E-step calculates the function $Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k)}) = \sum_{i=1}^{n} Q_i(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k)})$, where

$$
\begin{aligned}
Q_i(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k)}) &= \mathrm{E}\left\{\ell_c(\boldsymbol{\theta};\mathbf{y}_{i_c})|\boldsymbol{\theta}^{(k)},\mathbf{y}\right\} \qquad (11)\\
&\propto -\frac{3}{2}n_i\log\sigma - \frac{1}{2\sigma\tau_p^2}\Big[(\mathbf{y}_i - \mathbf{x}_i^\top\boldsymbol{\beta}_p)^\top \widehat{\mathbf{D}_i^{-1}}^{(k)}(\mathbf{y}_i - \mathbf{x}_i^\top\boldsymbol{\beta}_p)\\
&\quad -2(\mathbf{y}_i - \mathbf{x}_i^\top\boldsymbol{\beta}_p)\widehat{(\mathbf{D}_i^{-1}\mathbf{z}\mathbf{b})_i}^{(k)} + \mathrm{tr}\left\{\mathbf{z}_i\widehat{(\mathbf{b}\mathbf{b}^\top\mathbf{z}\mathbf{D}_i^{-1})_i}^{(k)}\right\}\\
&\quad -2\vartheta_p(\mathbf{y}_i - \mathbf{x}_i^\top\boldsymbol{\beta}_p)^\top\mathbf{1}_{n_i} + 2\vartheta_p(\mathbf{z}\widehat{\mathbf{b}}^{(k)})_i^\top\mathbf{1}_{n_i} + \frac{\tau_p^4}{4}\widehat{\mathbf{u}}_i^{(k)\top}\mathbf{1}_{n_i}\Big]\\
&\quad -\frac{1}{2}log|\boldsymbol{\Psi}| - \frac{1}{2}\mathrm{tr}\left\{\widehat{(\mathbf{b}\mathbf{b}^\top)_i}^{(k)}\boldsymbol{\Psi}^{-1}\right\},
\end{aligned}
$$

where $\mathrm{tr}(\mathbf{A})$ indicates the trace of matrix $\mathbf{A}$ and $\mathbf{1}_p$ is the vector of ones of dimension p. The calculation of these functions require expressions for

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

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

$$
\mathrm{E}[g(\mathbf{b}_i,\mathbf{u}_i)\,|\boldsymbol{\theta}^{(k)},\mathbf{y}_i] \approx \frac{1}{m}\sum_{\ell=1}^{m} g(\mathbf{b}_i^{(\ell,k)},\mathbf{u}_i^{(\ell,k)}), \qquad (12)
$$

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 $f(\mathbf{b}_i,\mathbf{u}_i|\boldsymbol{\theta}^{(k)},\mathbf{y}_i)$. A Gibbs Sampler can be easily implemented (see supplementary material) given that the two full conditional distributions $f(\mathbf{b}_i|\boldsymbol{\theta}^{(k)},\mathbf{u}_i,\mathbf{y}_i)$ and $f(\mathbf{u}_i|\boldsymbol{\theta}^{(k)},\mathbf{b}_i,\mathbf{y}_i)$ are known. However, using known properties of conditional expectations, the expected value in (12) can be more accurately approximated as

$$
\begin{aligned}
\mathrm{E}_{\mathbf{b}_i,\mathbf{u}_i}[g(\mathbf{b}_i,\mathbf{u}_i)|\boldsymbol{\theta}^{(k)},\mathbf{y}_i] &= \mathrm{E}_{\mathbf{b}_i}[\mathrm{E}_{\mathbf{u}_i}[g(\mathbf{b}_i,\mathbf{u}_i)|\boldsymbol{\theta}^{(k)},\mathbf{b}_i,\mathbf{y}_i]|\mathbf{y}_i]\\
&\approx \frac{1}{m}\sum_{\ell=1}^{m}\mathrm{E}_{\mathbf{u}_i}[g(\mathbf{b}_i^{(\ell,k)},\mathbf{u}_i)|\boldsymbol{\theta}^{(k)},\mathbf{b}_i^{(\ell,k)},\mathbf{y}_i], \qquad (13)
\end{aligned}
$$

where $\mathbf{b}^{(\ell,k)}$ is a sample from the conditional density $f(\mathbf{b}_i|\boldsymbol{\theta}^{(k)},\mathbf{y}_i)$. Note that (13) is a more accurate approximation as it only depends of one MC approximation instead two as needed in (12).

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 $\mathbf{u}_i|\mathbf{y}_i,\mathbf{b}_i = [\; u_{i1}|y_{i1},\mathbf{b}_i, \;\; u_{i2}|y_{i2},\mathbf{b}_i, \;\; \cdots \;\; ,u_{in_i}|y_{in_i},\mathbf{b}_i \;]^\top$, 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

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

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

$$\chi_{ij} = \frac{\frac{|y_{ij}-\mathbf{x}_{ij}^{\top}\boldsymbol{\beta}_p-\mathbf{z}_{ij}^{\top}\mathbf{b}_i|}{}}{\tau_p\sqrt{\sigma}} \quad \text{and} \quad \psi = \frac{\tau_p}{2\sqrt{\sigma}} \tag{14}$$

From (5), and after generating samples from $f(\mathbf{b}_i|\boldsymbol{\theta}^{(k)},\mathbf{y}_i)$ (see Subsection 3.3), the conditional expectation $\mathrm{E}_{\mathbf{u}_i}[\cdot|\boldsymbol{\theta},\mathbf{b}_i,\mathbf{y}_i]$ in (13) can be computed analytically. Finally, the proposed MCEM algorithm for estimating the parameters of the QR-LMM 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 3.3.

- **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:

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

where $N = \sum_{i=1}^{n} n_i$ and expressions $\mathscr{E}(\mathbf{u}_i)^{(\ell,k)}$ and $\mathscr{E}(\mathbf{D}_i^{-1})^{(\ell,k)}$ are defined in Appendix A.2 of the Supplementary Material. 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 3.3.

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 $i = 1,\ldots,n$;

- **Simulation step:** Draw $\mathbf{b}_i^{(\ell,k)}$, $\ell = 1,\ldots,m$, from $f(\mathbf{b}_i|\boldsymbol{\theta}^{(k)},\mathbf{y}_i)$, for $m \leq 20$.
- **Stochastic approximation:** Update the MC approximations for the conditional expectations by their stochastic approximations, given by

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

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

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_p^{(k+1)} &= \left[\sum_{i=1}^{n} S_{1,i}^{(k)}\right]^{-1} \sum_{i=1}^{n} S_{2,i}^{(k)}, \\
\widehat{\sigma}^{(k+1)} &= \frac{1}{3N\tau_p^2} \sum_{i=1}^{n} S_{3,i}^{(k)}, \\
\widehat{\boldsymbol{\Psi}}^{(k+1)} &= \frac{1}{n} \sum_{i=1}^{n} S_{4,i}^{(k)}. \qquad (15)
\end{aligned}$$

Given a set of suitable initial values $\widehat{\boldsymbol{\theta}}^{(0)}$ (see Appendix A.1 of the Supplementary Material), the SAEM iterates till convergence at iteration k, if $\max_i \left\{ \frac{|\widehat{\theta}_i^{(k+1)} - \widehat{\theta}_i^{(k)}|}{|\widehat{\theta}_i^{(k)}| + \delta_1} \right\} < \delta_2$, the stopping criterion, is satisfied for three consecutive times, where δ_1 and δ_2 are pre-established small values. This consecutive evaluation avoids a fake convergence produced by an unlucky Monte Carlo simulation. As suggested by Searle et al. (1992) (page. 269), we use $\delta_1 = 0.001$ and $\delta_2 = 0.0001$. This proposed criterion will need an extremely large number of iterations (more than usual) in order to detect parameter convergence 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 above and the convergence will not be attained even though the likelihood was maximized with few iterations. As proposed by Booth and Hobert (1999), we also use a second convergence criteria defined by $\max_i \left\{ \frac{|\widehat{\theta}_i^{(k+1)} - \widehat{\theta}_i^{(k)}|}{\sqrt{\widehat{\mathrm{var}}(\theta_i^{(k)}) + \delta_1}} \right\} < \delta_2$, where the parameter estimates change relative to their standard errors leading to a convergence detection even for bounded parameters. Once again, δ_1 and δ_2 are some small pre-assigned values, not necessarily equal to the ones in the previous criterion. Based on simulation results, we fix $\delta_1 = 0.0001$ and $\delta_2 = 0.0002$. This stopping criteria is similar to the one proposed by Bates and 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 $f(\mathbf{b}_i|\theta^{(k)},\mathbf{y}_i)$. 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 (7)-(8) is given by

$$\ell_o(\boldsymbol{\theta}|\mathbf{y})=\sum_{i=1}^{n}\log f(\mathbf{y}_i|\boldsymbol{\theta}))=\sum_{i=1}^{n}\log\int_{\mathbb{R}^q} f(\mathbf{y}_i|\mathbf{b}_i;\boldsymbol{\theta})\,f(\mathbf{b}_i;\boldsymbol{\theta})\,d\mathbf{b}_i, \qquad (16)$$

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 (16) 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

$$\ell_o(\boldsymbol{\theta}|\mathbf{y}) \approx \sum_{i=1}^{n} \log \left\{ \frac{1}{m} \sum_{\ell=1}^{m} \left[\prod_{j=1}^{n_i} [f(y_{ij}|\mathbf{b}_i^{*(\ell)};\boldsymbol{\theta})] \frac{f(\mathbf{b}_i^{*(\ell)};\boldsymbol{\theta})}{\widehat{f}(\mathbf{b}_i^{*(\ell)};\boldsymbol{\theta})} \right] \right\}, \tag{17}$$

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 3.3, 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.

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

$$\mathbf{I}_e(\boldsymbol{\theta}|\mathbf{y}) = \sum_{i=1}^{n} \mathbf{s}(\mathbf{y}_i|\boldsymbol{\theta})\, \mathbf{s}^\top(\mathbf{y}_i|\widehat{\boldsymbol{\theta}}) - \frac{1}{n} \mathbf{S}(\mathbf{y}|\boldsymbol{\theta})\, \mathbf{S}^\top(\mathbf{y}|\boldsymbol{\theta}), \tag{18}$$

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 $\boldsymbol{\nabla}_o(\hat{\boldsymbol{\theta}}) = \mathbf{0}$, equation (18) takes the simple form

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

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

$$\mathrm{s}(\mathbf{y}_i|\boldsymbol{\theta})^{(k)} = \mathbf{s}(\mathbf{y}_i|\boldsymbol{\theta})^{(k-1)} + \delta_k \left[\frac{1}{m} \sum_{\ell=1}^{m} \mathbf{s}(\mathbf{y}_i, \mathbf{q}^{(k,\ell)};\boldsymbol{\theta}^{(k)}) - \mathbf{s}(\mathbf{y}_i|\boldsymbol{\theta})^{(k-1)} \right], \tag{20}$$

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.3 of the Supplementary Material.

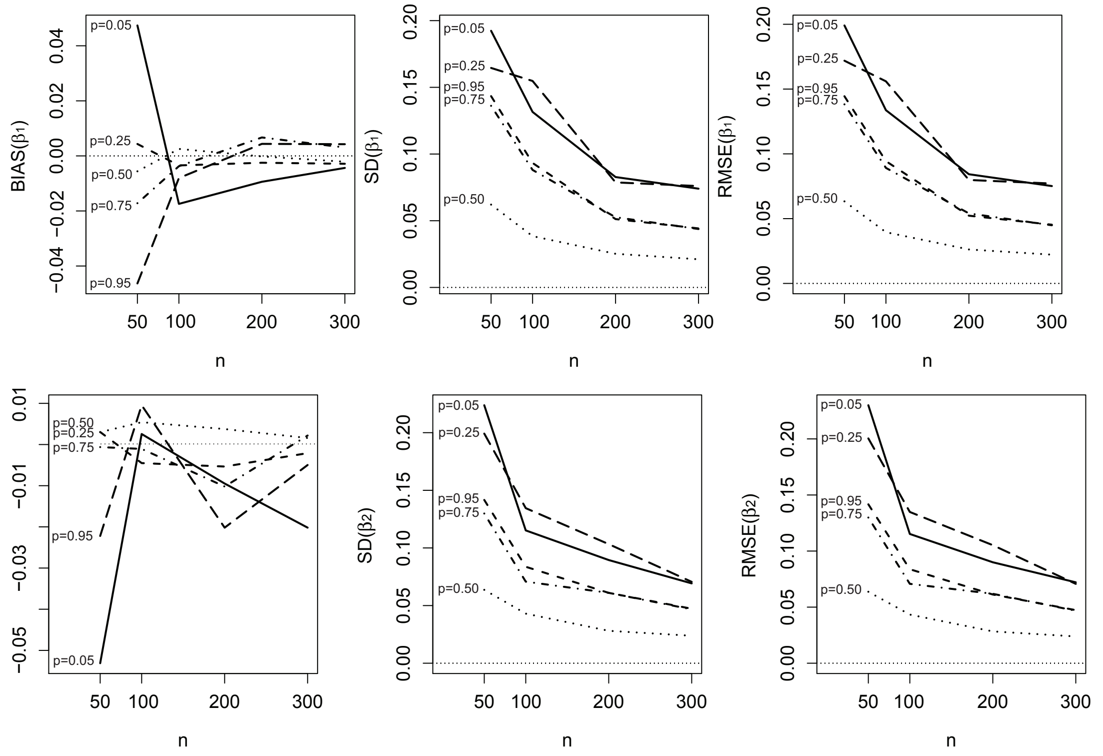

Figure 2: Bias, Standard Deviation and RMSE for β_1 (upper panel) and β_2 (lower panel) for varying sample sizes over the quantiles $p = 0.05, 0.10, 0.50, 0.90, 0.95$.

5 Simulation studies

In this section, the finite sample performance of the proposed algorithm and its performance comparison with the Geraci and Bottai (2014) method is evaluated via simulation studies. These computational procedures were implemented using the `R` software (R Core Team, 2014). In particular, we consider the following linear mixed model:

$$y_{ij} = \mathbf{x}_{ij}^{\top}\boldsymbol{\beta} + \mathbf{z}_{ij}\mathbf{b}_i + \varepsilon_{ij},\ i = 1,\ldots,n,\ j = 1,\ldots,3, \tag{21}$$

where the goal is to estimate the fixed effects parameters $\boldsymbol{\beta}$ for a grid of percentiles $p = \{0.05, 0.10, 0.50, 0.90, 0.95\}$. We simulated a 3×3 design matrix $\mathbf{x}_{ij}^{\top}$ for the fixed effects $\boldsymbol{\beta}$, where the first column corresponds to the intercept and the other columns generated from a $N_2(\mathbf{0}, \mathbf{I}_2)$ density, for all $i = 1,\ldots,n$. We also simulated a 3×2 design matrix associated with the random effects, with the columns distributed as $N_2(\mathbf{0}, \mathbf{I}_2)$. The fixed effects parameters were chosen as $\beta_1 = 0.8$, $\beta_2 = 0.5$ and $\beta_3 = 1$, $\sigma = 0.20$, and the matrix $\boldsymbol{\Psi}$ with elements $\Psi_{11} = 0.8$, $\Psi_{12} = 0.5$ and $\Psi_{22} = 1$. For varying sample sizes of $n = 50, 100, 200$ and 300, we generate 100 data samples for each scenario. In addition, we also choose $m = 20$, $c = 0.2$ and $W = 500$.

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 $\text{MC-Sd}(\widehat{\theta}_i) = \sqrt{\frac{1}{99}\sum_{j=1}^{100}\left(\widehat{\theta}_i^{(j)} - \overline{\widehat{\theta}_i}\right)^2}$, $\text{Bias}(\widehat{\theta}_i) = \overline{\widehat{\theta}_i} - \theta_i$, and $\text{RMSE}(\widehat{\theta}_i) = \sqrt{\text{MC-Sd}^2(\widehat{\theta}_i) + \text{Bias}^2(\widehat{\theta}_i)}$, where $\overline{\widehat{\theta}_i} = \frac{1}{100}\sum_{j=1}^{100}\widehat{\theta}_i^{(j)}$ and $\theta_i{}^{(j)}$ is the estimate of θ_i from the j-th sample, $j = 1\ldots 100$. In addition, we also computed the average of the standard deviations (IM-Sd) obtained via the observed information matrix derived in Subsection 4.2 and the 95%

Table 1: Monte Carlo standard deviation (MC-Sd), mean standard deviation (IM-Sd) and Monte Carlo coverage probability (MC-CP) estimates of the fixed effects β_1 and β_2 from fitting the QR-LMM under various quantiles for sample size $n = 100$.

	β_1			β_2		
Quantile (%)	MC-Sd	IM-Sd	MC-CP	MC-Sd	IM-Sd	MC-CP
5	0.073	0.060	90	0.067	0.059	90
10	0.045	0.044	95	0.047	0.044	96
50	0.022	0.024	97	0.024	0.025	96
90	0.045	0.045	92	0.047	0.044	96
95	0.060	0.056	88	0.071	0.056	83

coverage probability (MC-CP) as $\text{CP}(\hat{\theta}_i) = \frac{1}{100}\sum_{j=1}^{100} I(\theta_i \in [\hat{\theta}_{i,LCL}, \hat{\theta}_{i,UCL}])$, where I is the indicator function such that θ_i lies in the interval $[\hat{\theta}_{i,LCL}, \hat{\theta}_{i,UCL}]$, with $\hat{\theta}_{i,LCL}$ and $\hat{\theta}_{i,UCL}$ as the estimated lower and upper bounds of the 95% CIs, respectively.

Table 2: Simulation 1: Root Mean Squared Error (RMSE) for the fixed effects β_0, β_1, β_2 and the nuisance parameter σ, obtained after fitting the QRLMM and the Geraci (2014) model to simulated data under various settings of quantiles and sample sizes.

		RMSE							
		β_0		β_1		β_2		σ	
Quantile (%)	n	SAEM	Geraci	SAEM	Geraci	SAEM	Geraci	SAEM	Geraci
5	50	0.249	0.622	0.199	0.311	0.230	0.296	0.024	0.046
	100	0.209	0.496	0.134	0.180	0.115	0.165	0.017	0.037
	200	0.195	0.303	0.084	0.099	0.090	0.137	0.017	0.029
	300	0.163	0.345	0.075	0.100	0.072	0.101	0.012	0.031
10	50	0.159	0.382	0.144	0.187	0.142	0.201	0.023	0.048
	100	0.112	0.355	0.094	0.117	0.084	0.130	0.019	0.048
	200	0.082	0.231	0.052	0.087	0.061	0.081	0.017	0.036
	300	0.073	0.223	0.045	0.072	0.047	0.076	0.011	0.034
50	50	0.063	0.107	0.063	0.090	0.064	0.102	0.025	0.174
	100	0.042	0.052	0.040	0.056	0.043	0.070	0.021	0.196
	200	0.027	0.053	0.026	0.048	0.028	0.039	0.016	0.164
	300	0.024	0.034	0.022	0.022	0.024	0.040	0.012	0.180
90	50	0.160	0.389	0.138	0.159	0.130	0.177	0.025	0.050
	100	0.102	0.394	0.089	0.100	0.071	0.126	0.019	0.051
	200	0.085	0.240	0.054	0.097	0.062	0.078	0.014	0.038
	300	0.065	0.276	0.045	0.066	0.047	0.064	0.011	0.038
95	50	0.255	0.552	0.172	0.255	0.200	0.243	0.020	0.040
	100	0.233	0.470	0.156	0.169	0.135	0.161	0.020	0.036
	200	0.146	0.423	0.080	0.160	0.105	0.106	0.015	0.038
	300	0.157	0.468	0.077	0.113	0.071	0.061	0.014	0.036

The results are summarized in Figure 2. We observe that the *Bias*, *SD* and *RMSE* for the regression parameters β_1 and β_2 tends to approach zero with increasing sample size (n), revealing that the ML estimates obtained via the proposed SAEM algorithm are conformable to the expected asymptotic properties. In addition, Table 1 presents the IM Sd, MC-Sd and MC-CP for β_1 and β_2 across various quantiles. The estimates of MC-Sd and IM-Sd are very close, hence we can infer that the asymptotic approximation of the parameter standard errors are reliable. Furthermore, as expected, we observe that the MC-CP remains lower for extreme quantiles.

Finally, we compare the performance of SAEM algorithm with the approximate method proposed by Geraci (2014). The Geraci's algorithm can be implemented using the `R` package `lqmm()`. The results are presented in Table 2 and Figure B.1 (Supplementary Material). We observe that the RMSE from the proposed SAEM algorithm are lower than Geraci method across all scenarios, with the differences considerably higher for the extreme quantiles. Finally, Figure B.2 (Supplementary Material) that compares the differences in SD between the two methods for fixed effects β_1 and β_2 at specified quantiles reveals that the SD are mostly smaller for the SAEM method. Thus, we conclude that the SAEM algorithm produces more precise estimates.

6 Applications

In this section, we illustrate the application of our method to two interesting longitudinal datasets from the literature via our developed `R` package `qrLMM`, currently available for free download from the `R CRAN` (Comprehensive R Archive Network).

6.1 Cholesterol data

The Framingham cholesterol study generated a benchmark dataset (Zhang and Davidian, 2001) for longitudinal analysis to examine the role of serum cholesterol as a risk factor for the evolution of cardiovascular disease. We analyze this dataset with the aim of explaining the full conditional distribution of the serum cholesterol as a function of a set of covariates of interest via modelling a grid of response quantiles. We fit a LMM model to the data as specified by

$$Y_{ij} = \beta_0 + \beta_1 \text{gender}_i + \beta_2 \text{age}_i + b_{0i} + b_{1i} t_{ij} + \varepsilon_{ij}, \tag{22}$$

where Y_{ij} is the cholesterol level (divided by 100) at the jth time point for the ith subject, $t_{ij} = (\tau - 5)/10$ where τ is the time measured in years from the start of the study, age denotes the subject's baseline age, gender is the dichotomous gender (0=female, 1=male), b_{0i} and b_{1i} the random intercept and slope, respectively, for subject i, and ε_{ij} the measurement error term, for 200 randomly selected subjects.

After fitting the QR-LMM over the grid $p = \{0.05, 0.10, \ldots, 0.95\}$, we present a graphical summary of the results in Figure 3. The figure displays the 95% confidence band for the fixed effects parameters $\beta_0, \beta_1, \beta_2$, and for the nuisance parameter σ. The solid lines represent the $Q_{0.025}$ and $Q_{0.975}$ percentiles, obtained from the estimated standard errors defined in Subsection 4.2. The figure reveals that the effect of gender and age become more prominent with increasing conditional quantiles (p). In addition, although age exhibits a positive influence on the cholesterol level across all quantiles, the confidence band for gender includes 0 across all quantiles, and hence its effect is non-significant. The estimated nuisance parameter σ is symmetric about $p = 0.5$, taking its maximum value at that point and decreasing for the extreme quantiles. Figure B.3 (Supplementary Material) plots the fitted regression lines for the quantiles $0.10, 0.25, 0.50(\text{mean}), 0.75$ and 0.90 by gender. From this figure, it is clear how the extreme quantiles capture the full data variability and detect some atypical observations. The intercept of the quantile functions look very similar for both panels because of the non-significance of gender.

6.2 Orthodontic distance growth data

A second application was developed using a data set form a longitudinal orthodontic study (Potthoff and Roy, 1964; Pinheiro et al., 2001) performed at the University of North Carolina Dental School. Here, researchers measured the distance between the pituitary and the pterygomaxillary fissure (two points that are easily identified on x-ray exposures of the side of the head) for 27 children (16 boys and 11 girls) every two years from age 8 until age 14. Similar to Application 1, we fit the following LMM to the data:

$$Y_{ij} = \beta_0 + \beta_1 \text{gender}_i + \beta_2 t_{ij} + b_{0i} + b_{1i} t_{ij} + \varepsilon_{ij}, \tag{23}$$

p0
where Y_{ij} is the distance between the pituitary and the pterygomaxillary fissure (in mm) at the jth time for the ith child, t_{ij} is the child's age at time j taking values 8, 10, 12, and 14 years, gender is a dichotomous variable (0=female, 1=male) for child i and ε_{ij} the random measurement error term. Initial exploratory plots for 10 random children in the left panel of Figure B.4 (Supplementary Material) suggest an increasing distance with respect to age. The individual profiles by gender (right panel) show differences between distances for boys and girls (distance for boys greater than those for girls), and hence we could expect a significant gender effect. Once again, after fitting the QR-LMM over the grid $p = \{0.05, 0.10, \ldots, 0.95\}$, the point estimates and associated 95% confidence bands for model parameters are presented in Figure 4. From the figure, we infer that the effect of gender and age are significant across all quantiles, with their effect increasing for higher conditional quantiles. Effect of Age is always positive across all quantiles, with a higher effect at the two extremes. σ behaves the same as in Application 1. Figure B.5 (Supplementary Material) plots the fitted regression lines for the quantiles $0.10, 0.25, 0.50, 0.75$ and 0.90, overlayed with the individual profiles (gray solid lines), by gender. These fits cap-

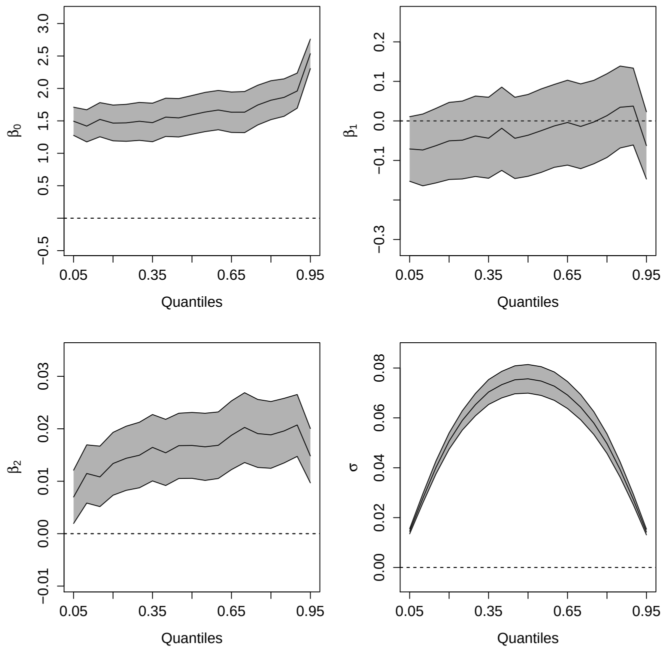

Figure 3: Point estimates (center solid line) and 95% confidence intervals for model parameters after fitting the QR-LMM using the `qrLMM` package to the Cholesterol data across various quantiles. The interpolated curves are spline-smoothed.

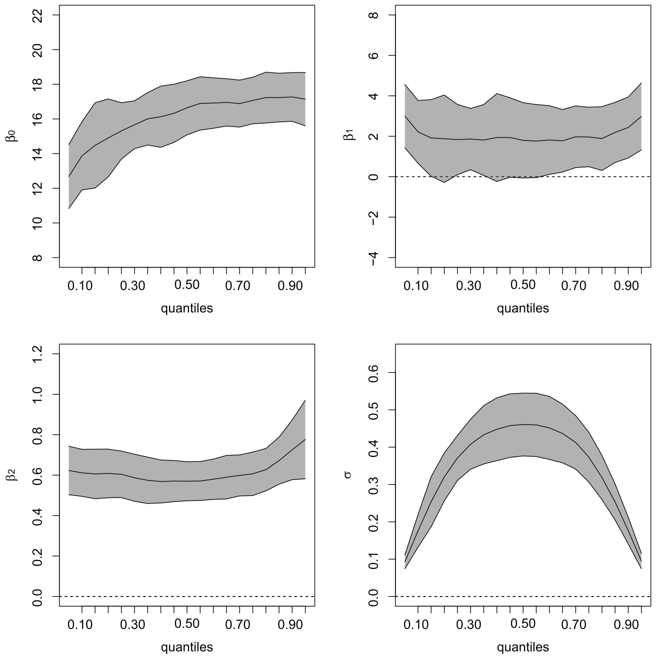

Figure 4: Point estimates (center solid line) and 95% confidence intervals for model parameters after fitting the QR-LMM using the `qrLMM` package to the orthodontic growth distance data across various quantiles. The interpolated curves are spline-smoothed.

ture the variability of the individual profiles, and also differ by gender due to its significance in the model. The `R` package also produces graphical summaries of point estimates and confidence intervals (95% by default) across various quantiles, as presented in Figures 3 and 4. Trace plots showing convergence of these estimates are presented in Figure B.6 (Supplementary Material). For example, for the 75th quantile, we can confirm that the convergence parameters for the SAEM algorithm ($M = 10$, $c = 0.25$ and $W = 300$) has been set adequately leading to a quick convergence in distribution within the first 75 iterations, and then converging almost surely to a local maxima in a total of 300 iterations. Sample output from the `qrLMM` package is provided in Appendix C of the Supplementary Material.

7 Conclusions

In this paper, we developed a likelihood-based inference for QR-LMM with the likelihood function based on the ALD. The ALD presents a convenient framework for the implementation of the SAEM algorithm leading to the exact ML estimation of the parameters. The methodology is illustrated via application to two longitudinal clinical datasets. We believe this paper is the first attempt for exact ML estimation in the context of QR-LMMs, and thus provides an improvement over the Geraci and Bottai (2014) method. The methods developed are readily implementable via the `R` package `qrLMM()`.

Although the QR-LMM considered here has shown great flexibility to quantify the entire conditional distribution of the outcome variable, its robustness against outliers can be seriously affected by the presence of skewness and thick-tails. Recently, Lachos et al. (2010) proposed a remedy to accommodate these using scale mixtures of skew-normal distributions in the random effects. We conjecture that methodology can be transferred to the QR-LMM framework, and should yield satisfactory results at the expense of additional complexity in implementation. An in-depth investigation of such extension is beyond the scope of the present paper, but certainly an interesting topic for future research.

APPENDIX A Some results on SAEM implementation

A.1 A Gibbs Sampler Algorithm

In order to draw a sample from $f(\mathbf{b}_i, \mathbf{u}_i|\mathbf{y}_i)$ we can use the Gibbs Sampler, an Markov chain Monte Carlo (MCMC) algorithm proposed by (Casella and George, 1992) for obtaining a sequence of observations which are approximated from the joint probability distribution of two or several random variables just using their full conditional distributions. Computing the full conditional distributions $f(\mathbf{b}_i|\mathbf{u}_i, \mathbf{y}_i)$ and $f(\mathbf{u}_i|\mathbf{b}_i, \mathbf{y}_i)$, we have for the first one that

$$\begin{aligned} f(\mathbf{b}_i|\mathbf{y}_i, \mathbf{u}_i) &\propto f(\mathbf{y}_i|\mathbf{b}_i, \mathbf{u}_i)\, f(\mathbf{b}_i), \\ &\propto \phi_{n_i}\left(\mathbf{y}_i|\, \mathbf{X}_i^\top \boldsymbol{\beta}_p + \mathbf{Z}_i\mathbf{b}_i + \vartheta_p\mathbf{u}_i, \sigma\tau_p^2 D(\mathbf{u}_i)\right) \times \phi_q(\mathbf{b}_i|\mathbf{0}, \boldsymbol{\Psi}) \end{aligned} \tag{24}$$

so we have a product of multivariate normal densities which solution is based in the next lemma:

Lemma 1. *Simplifying the notation above it follows that*

$$\phi_n(\mathbf{y}|\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b}, \boldsymbol{\Omega})\phi_q(\mathbf{b}|\mathbf{0}, \boldsymbol{\Psi}) = \phi_n(\mathbf{y}|\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma})\phi_q(\mathbf{b}|\boldsymbol{\mu}_1(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \boldsymbol{\Lambda}) \tag{25}$$

where

$$\boldsymbol{\mu}_1 = \boldsymbol{\Lambda}\mathbf{Z}^T\boldsymbol{\Omega}^{-1}, \quad \boldsymbol{\Sigma} = \boldsymbol{\Omega} + \mathbf{Z}\boldsymbol{\Psi}\mathbf{Z}^T, \quad \boldsymbol{\Lambda} = (\boldsymbol{\Psi}^{-1} + \mathbf{Z}^T\boldsymbol{\Omega}^{-1}\mathbf{Z})^{-1}. \tag{26}$$

Due the equation (25) from the lemma 2 it leads us to

$$\begin{aligned} f(\mathbf{b}_i|\mathbf{y}_i, \mathbf{u}_i) \propto\ & \phi_{n_i}\left(\mathbf{y}_i|\ \mathbf{X}_i^\top \boldsymbol{\beta}_p + \vartheta_p\mathbf{u}_i, \sigma\tau_p^2 D(\mathbf{u}_i) + \mathbf{Z}_i\Psi\mathbf{Z}_i^\top\right) \times \\ & \phi_q\left(\mathbf{b}_i|\, \boldsymbol{\Lambda}_i\mathbf{Z_i}^\top\left(\sigma\tau_p^2 D(\mathbf{u}_i)\right)^{-1}\left(\mathbf{y}_i - \mathbf{X}_i^\top\boldsymbol{\beta}_p - \vartheta_p\mathbf{u}_i\right), \boldsymbol{\Lambda}_i\right) \end{aligned}$$

where $\boldsymbol{\Lambda}_i = \left(\boldsymbol{\Psi}^{-1} + \sigma\tau_p^2\mathbf{Z}_i^\top D(\mathbf{u}_i)\mathbf{Z}_i\right)^{-1}$. Then dropping the first term of the product by proportionality it's easy to see that $\mathbf{b}_i|\mathbf{y}_i, \mathbf{u}_i \sim N_q\left(\boldsymbol{\Lambda}_i\mathbf{Z}_i^\top\left(\sigma\tau_p^2 D(\mathbf{u}_i)\right)^{-1}\left(\mathbf{y}_i - \mathbf{X}_i^\top\boldsymbol{\beta}_p - \vartheta_p\mathbf{u}_i\right), \boldsymbol{\Lambda}_i\right)$.

On other hand, for the full conditional distribution $f(\mathbf{u}_i|\mathbf{y}_i, \mathbf{b}_i)$ note that the vector $\mathbf{u}_i|\mathbf{y}_i, \mathbf{b}_i$ can be constructed as $\mathbf{u}_i|\mathbf{y}_i, \mathbf{b}_i = \left[\begin{array}{cccc} u_{i1}|y_{i1}, \mathbf{b}_i & u_{i2}|y_{i2}, \mathbf{b}_i & \cdots & u_{in_i}|y_{in_i}, \mathbf{b}_i \end{array}\right]^\top$ given that $u_{ij}\big|y_{ij}, \mathbf{b}_i \perp u_{ik}|y_{ik}, \mathbf{b}_i$ for all $j, k = 1, 2, \ldots, n_i$ and $j \neq k$. So, the univariate distribution of the $f(u_{ij}|y_{ij}, \mathbf{b}_i)$ is proportional to the product of $f(y_{ij}|\mathbf{b}_i, u_{ij})$ and $f(u_{ij})$, a Normal and a Exponential distribution, that is

$$f(u_{ij}|y_{ij}, \mathbf{b}_i) \propto \phi(y_{ij}\big|\mathbf{X}_{ij}^\top\boldsymbol{\beta}_p + \mathbf{Z}_{ij}^\top\mathbf{b}_i + \vartheta_p u_{ij},\, \sigma\tau_p^2 u_{ij}) \times G_{U_{ij}}(1, \sigma),$$

then the Lemma 1 leads us that $u_{ij}|y_{ij}, \mathbf{b}_i \sim GIG(\frac{1}{2}, \chi_{ij}, \psi)$, where $\chi_{ij} = \dfrac{\left|y_{ij} - \mathbf{X}_{ij}^\top\boldsymbol{\beta}_p - \mathbf{Z}_{ij}^\top\mathbf{b}_i\right|}{\tau_p\sqrt{\sigma}}$ and $\psi = \dfrac{\tau_p}{2\sqrt{\sigma}}$.

In resume, the Gibbs Sampler proceeds as follow:

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

(1) Start with suitable initial values $(\mathbf{b}_i^{(0,k)}, \mathbf{u}_i^{(0,k)})$

(2) Draw $\mathbf{b}_i^{(1,k)} | \mathbf{y}_i, \mathbf{u}_i^{(0,k)} \sim N_q \left(\boldsymbol{\Lambda}_i^{(k)} \mathbf{Z}_i^\top \left(\sigma^{(k)} \tau_p^2 D(\mathbf{u}_i^{(0,k)}) \right)^{-1} \left(\mathbf{y}_i - \mathbf{X}_i^\top \boldsymbol{\beta}_p^{(k)} - \vartheta_p \mathbf{u}_i^{(0,k)} \right), \boldsymbol{\Lambda}_i^{(k)} \right)$

(3) Draw $u_{ij}^{(1,k)} | y_{ij}, \mathbf{b}_i^{(1,k)} \sim GIG \left(\frac{1}{2}, \frac{\left| y_{ij} - \mathbf{X}_{ij}^\top \boldsymbol{\beta}_p^{(k)} - \mathbf{Z}_{ij}^\top \mathbf{b}_i^{(1,k)} \right|}{\tau_p \sqrt{\sigma^{(k)}}}, \frac{\tau_p}{2\sqrt{\sigma^{(k)}}} \right)$ for all $j = 1, 2, \ldots, n_i$

(4) Construct $\mathbf{u}_i^{(1,k)} | \mathbf{y}_i, \mathbf{b}_i^{(1,k)}$ as $\left[\begin{array}{cccc} u_{i1}^{(1,k)} | y_{i1}, \mathbf{b}_i^{(1,k)} & u_{i2}^{(1,k)} | y_{i2}, \mathbf{b}_i^{(1,k)} & \cdots & u_{in_i}^{(1,k)} | y_{in_i}, \mathbf{b}_i^{(1,k)} \end{array} \right]^\top$

(5) Repeat the steps 2-4 until draw m samples $\left(\mathbf{b}_i^{(1,k)}, \mathbf{u}_i^{(1,k)} \right), \left(\mathbf{b}_i^{(2,k)}, \mathbf{u}_i^{(2,k)} \right), \ldots, \left(\mathbf{b}_i^{(m,k)}, \mathbf{u}_i^{(m,k)} \right)$ from $\mathbf{b}_i, \mathbf{u}_i | \theta^{(k)}, \mathbf{y}_i$.

Note that for a given a iteration k and for all $i = 1, \ldots, n$, drawing from the conditional distribution of the vector $\mathbf{u}_i^{(l,k)} | \mathbf{y}_i, \mathbf{b}_i^{(l,k)}$implies to draw from the univariate conditional distributions $u_{ij}^{(k)} | y_{ij}, \mathbf{b}_i^{(k)}$ for all $j = 1, 2, \ldots, n_i$, so this construction results in a heavy computational algorithm.

A.2 Specification of initial values

It is well known that a smart choice of the initial values of ML estimates can assure a fast convergence of an algorithm to the global maxima solution for the respective likelihood. Obviating the random effects term, let $\mathbf{y}_i \sim ALD(\mathbf{x}_i^\top \boldsymbol{\beta}_p, \sigma, p)$. Next, considering the MLEs of $\boldsymbol{\beta}_p$ and σ as defined in **?**) for this model, we follow the steps below for the QR-LMM implementation:

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

$$\widehat{\boldsymbol{\beta}}_p^{(0)} = \underset{\beta_p \in \mathbb{R}^k}{\text{arg min}} \sum_{i=1}^{n} \rho_p(\mathbf{y}_i - \mathbf{x}_i^\top \boldsymbol{\beta}_p).$$

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

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

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

A.3 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 $\mathbf{u}_i | \mathbf{y}_i, \mathbf{b}_i = \left[\begin{array}{cccc} u_{i1} | y_{i1}, \mathbf{b}_i & u_{i2} | y_{i2}, \mathbf{b}_i & \cdots & u_{in_i} | y_{in_i}, \mathbf{b}_i \end{array} \right]^\top$. 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

$$\mathscr{E}(\mathbf{u}_i) = [\mathscr{E}(u_{i1}) \; \mathscr{E}(u_{i1}) \; \cdots \; \mathscr{E}(u_{in_i})]^\top \tag{27}$$

and

$$\mathscr{E}(\mathbf{D}_i^{-1}) = \mathrm{diag}(\mathscr{E}(\mathbf{u}_i^{-1})) = \begin{bmatrix} \mathscr{E}(u_{i1}^{-1}) & 0 & \dots & 0 \\ 0 & \mathscr{E}(u_{i2}^{-1}) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \mathscr{E}(u_{in_i}^{-1}) \end{bmatrix}. \tag{28}$$

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 (14). 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 (27)-(28) where

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

A.4 The empirical information matrix

In light of (10), the complete log-likelihood function can be rewritten as

$$\ell_{ci}(\boldsymbol{\theta}) = -\frac{3}{2} n_i \log \sigma - \frac{1}{2\sigma\tau_p^2} \zeta_i^\top \mathbf{D}_i^{-1} \zeta_i - \frac{1}{2} \log|\boldsymbol{\Psi}| - \frac{1}{2} \mathbf{b}_i^\top \boldsymbol{\Psi}^{-1} \mathbf{b}_i - \frac{1}{\sigma} \mathbf{u}_i^\top \mathbf{1}_{n_i} \tag{29}$$

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

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

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

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

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

$$\begin{aligned} \frac{\partial \ell_{ci}(\boldsymbol{\theta})}{\partial \alpha_j} &= \frac{\partial \boldsymbol{\Psi}}{\partial \alpha_j} \frac{\partial \ell_{ci}(\boldsymbol{\theta})}{\partial \boldsymbol{\Psi}} \\ &= \frac{\partial \boldsymbol{\Psi}}{\partial \alpha_j} \frac{1}{2} \mathrm{tr}\{\boldsymbol{\Psi}^{-1} (\mathbf{b}_i \mathbf{b}_i^\top - \boldsymbol{\Psi}) \boldsymbol{\Psi}^{-1}\}. \end{aligned} \tag{30}$$

where, using the fact that $\mathrm{tr}\{\mathbf{ABCD}\} = (\mathrm{vec}(\mathbf{A}^\top))^\top(\mathbf{D}^\top \otimes \mathbf{B})(\mathrm{vec}(\mathbf{C}))$, (30) can be rewritten as

$$\frac{\partial \ell_{ci}(\boldsymbol{\theta})}{\partial \alpha_j} = (\mathrm{vec}(\frac{\partial \boldsymbol{\Psi}}{\partial \alpha_j}^\top))^\top \frac{1}{2}(\boldsymbol{\Psi}^{-1} \otimes \boldsymbol{\Psi}^{-1})(\mathrm{vec}(\mathbf{b}_i\mathbf{b}_i^\top - \boldsymbol{\Psi})). \tag{31}$$

Let $\mathscr{D}_q$ be the elimination matrix **(?)** that transforms the vectorized $\boldsymbol{\Psi}$ (written as $\mathrm{vec}(\boldsymbol{\Psi})$) into its half-vectorized form vech($\boldsymbol{\Psi}$), such that $\mathscr{D}_q\mathrm{vec}(\boldsymbol{\Psi}) = \mathrm{vech}(\boldsymbol{\Psi})$. Using the fact that for all $j = 1,\ldots,\frac{1}{2}q(q+1)$, the vector $(\mathrm{vec}(\frac{\partial \boldsymbol{\Psi}}{\partial \alpha_j})^\top)^\top$ corresponds to the jth row of the elimination matrix $\mathscr{D}_q$, we can generalize the derivative in (31) for the vector of parameters $\boldsymbol{\alpha}$ as

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

Finally, at each iteration, we can compute the empirical information matrix (19) by approximating the score for the observed log-likelihood using the stochastic approximation given in (20).

APPENDIX B Figures

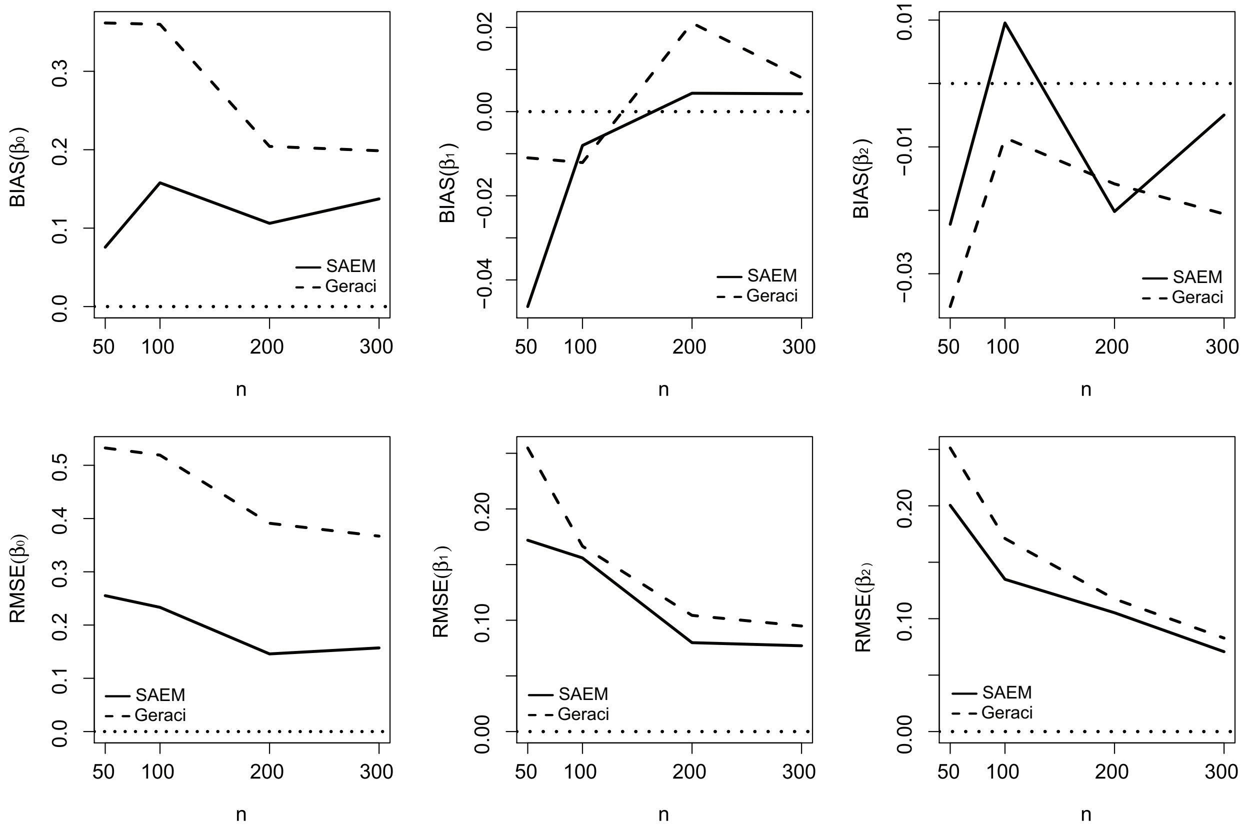

Figure B.1: Comparison of the Bias (upper row) and RMSE (lower row) at the 95-th quantile from fitting the QR-LMM and the Geraci (2014) model for the fixed effects β_0, β_1 and β_2.

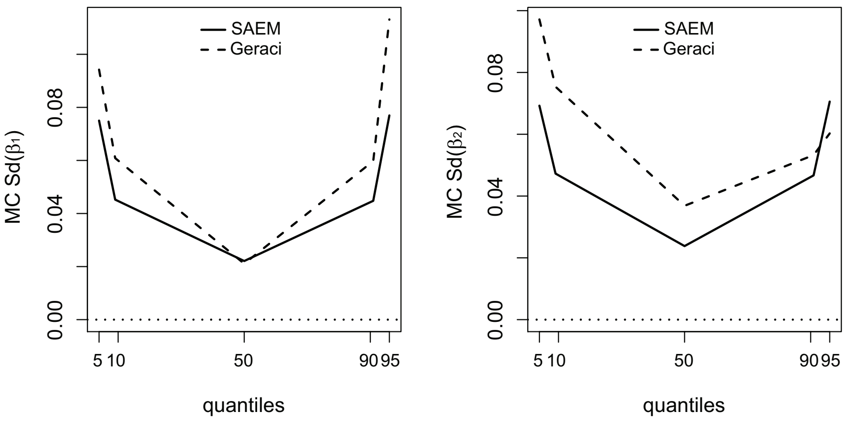

Figure B.2: Comparison of the Monte Carlo standard deviation for the estimatives of β_1 and β_2 obtained by the SAEM procedure and the Geraci (2014) algorithm for the set of quantiles 5, 10, 50, 90 and 95.

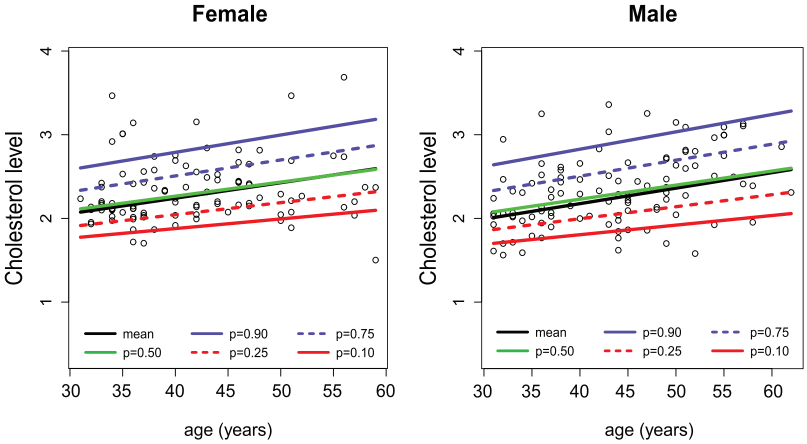

Figure B.3: Fitted mean regression overlayed with five different quantile regression lines for the Cholesterol data, by gender.

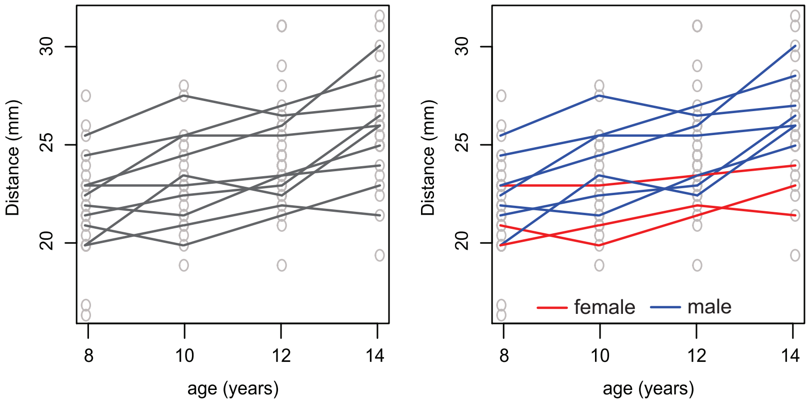

Figure B.4: Orthodontic distance growth data: Individual profiles for 10 random children (Panel a); Individual profiles for the same children, by gender (Panel b).

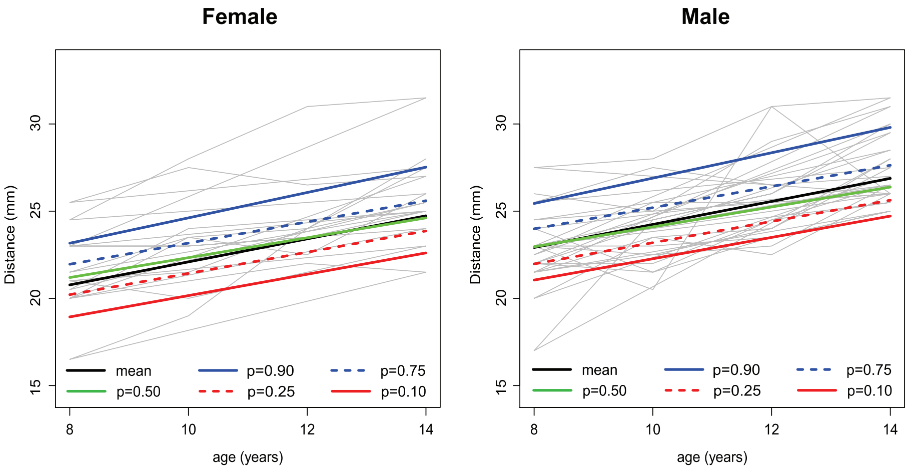

Figure B.5: Fitted mean regression overlayed with five different quantile regression lines for the Orthodontic distance growth data, by gender.

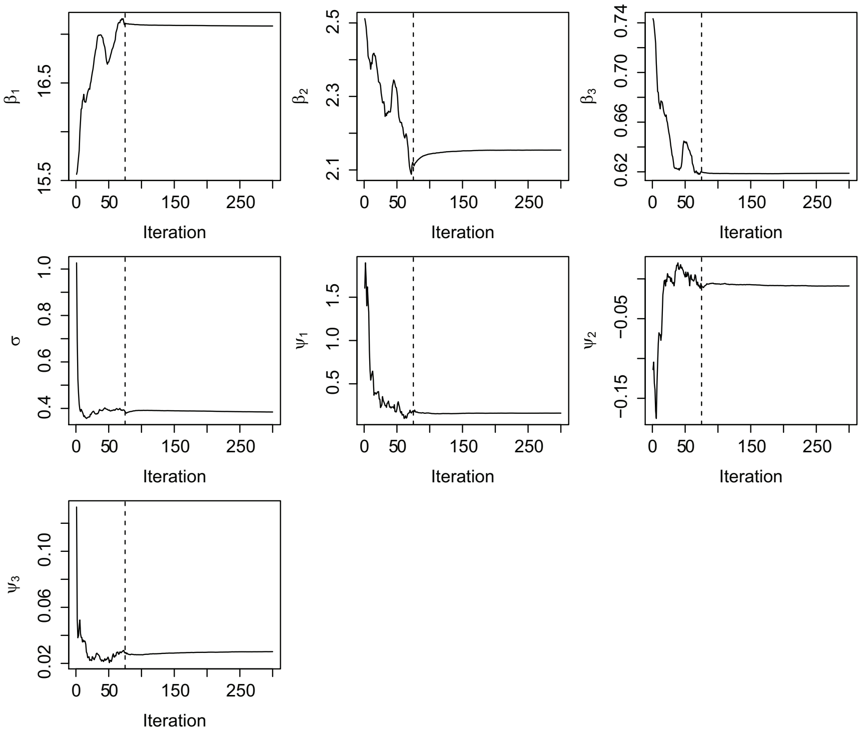

Figure B.6: Graphical summary of convergence for the fixed effect parameters, variance components of the random effects, and nuisance parameters, generated from the `qrLMM` package for the orthodontic distance growth data. The vertical dashed line delimits the beginning of the almost sure convergence, as defined by the cut-point parameter *c*.

APPENDIX C Sample output from `R` package `qrLMM()`

```
-------------------------------------------------
Quantile Regression for Linear Mixed Models
-------------------------------------------------
Quantile = 0.75
Subjects = 27 ; Observations = 108 ; Balanced = 4

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

       Estimate Std. Error  z value Pr(>|z|)
beta 1 17.08405    0.53524 31.91831        0
beta 2  2.15393    0.36929  5.83265        0
beta 3  0.61882    0.05807 10.65643        0

sigma = 0.38439

Random effects Variance-covariance matrix
         z1       z2
z1  0.16106 -0.00887
z2 -0.00887  0.02839

------------------------
Model selection criteria
------------------------
        Loglik     AIC     BIC     HQ
Value -216.454 446.907 465.682 454.52

-------
Details
-------
Convergence reached? = FALSE
Iterations = 300 / 300
Criteria = 0.00381
MC sample = 10
Cut point = 0.25
Processing time = 7.590584 mins
```

Acknowledgements

Bandyopadhyay acknowledges support from the US NIH grants R03DE021762 and R03DE023372. 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)

References

Allassonnière, S., E. Kuhn, A. Trouvé, 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. and N. Shephard (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, D. M. and D. G. Watts (1981). A Relative Off set Orthogonality Convergence Criterion for Nonlinear least Squares. *Technometrics 23*(2), 179–183.

Booth, J. G. and J. P. Hobert (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.

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

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

Galvao, A. F. and G. V. Montes-Rojas (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. (2014). Linear quantile mixed models: The lqmm package for laplace quantile regression. *Journal of Statistical Software 57*(13), 1–29.

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

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

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

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

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

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

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

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

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

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

Lipsitz, S. R., G. M. Fitzmaurice, G. Molenberghs, and L. P. Zhao (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) 51*(1), 127–138.

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

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

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

Pinheiro, J. C., C. Liu, and Y. N. Wu (2001). Efficient algorithms for robust estimation in linear mixed-effects models using the multivariate t distribution. *Journal of Computational and Graphical Statistics 10*(2), 249–276.

Potthoff, R. F. and S. Roy (1964). A generalized multivariate analysis of variance model useful especially for growth curve problems. *Biometrika 51*(3-4), 313–326.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

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

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

Wei, G. C. and M. A. Tanner (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, C. J. (1983). On the convergence properties of the EM algorithm. *The Annals of Statistics 11*(1), 95–103.

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

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

Zhang, D. and M. Davidian (2001). Linear mixed models with flexible distributions of random effects for longitudinal data. *Biometrics 57*(3), 795–802.