Finite Mixtures of Quantile and M-quantile regression models

Marco Alfò*   Nicola Salvati†   M. Giovanna Ranalli‡

November 10, 2016

Abstract

In this paper we define a finite mixture of quantile and M-quantile regression models for heterogeneous and /or for dependent/clustered data. Components of the finite mixture represent clusters of individuals with homogeneous values of model parameters. For its flexibility and ease of estimation, the proposed approaches can be extended to random coefficients with a higher dimension than the simple random intercept case. Estimation of model parameters is obtained through maximum likelihood, by implementing an EM-type algorithm. **The standard error estimates for model parameters are obtained using the inverse of the observed information matrix, derived through the Oakes (1999) formula in the M-quantile setting, and through nonparametric bootstrap in the quantile case.** We present a large scale simulation study to analyse the practical behaviour of the proposed model **and to evaluate the empirical performance of the proposed standard error estimates for model parameters. We considered** a variety of empirical settings in both the random intercept and the random coefficient case. The proposed modelling approaches are also applied to two well-known datasets which give further insights on their empirical behaviour.

Key words: Random effect models, nuisance parameters, nonparametric MLe, robust models, quantile regression, M-quantile regression, longitudinal data.

1 Introduction

Often, empirical applications entail dependent observations in the form of hierarchically structured data; these can derive, for example, through spatial, multilevel or longitudinal sample designs. When regression models for a response of interest are considered, the potential association between dependent observations should be taken into account

---

*Dipartimento di Scienze Statistiche, Sapienza Università di Roma, Italy, `marco.alfo@uniroma1.it`

†Dipartimento di Economia e Management, Università di Pisa, Italy, `salvati@ec.unipi.it`

‡Dipartimento di Scienze Politiche, Università degli Studi di Perugia, Italy `giovanna.ranalli@stat.unipg.it`

in order to provide valid and efficient inferences. The use of individual-specific random effects (or coefficients) helps us introduce a simple, potentially covariate-driven, structure of association between observations. Mixed effect models are now widely used in this context, and several estimation algorithms are available to users. Individual-specific random effects (coefficients) may also help us describe the influence of omitted covariates on parameter estimates for observed covariates. In this context, the random effects (coefficients) are often thought of as representing (individual-specific) unobserved heterogeneity, while observed covariates describe the observed counterpart.

Only recently random effects have been used to model dependence in quantile regression models for longitudinal data, by either using a time-constant dependence, as in Geraci and Bottai (2007, 2014) or adopting a time-varying structure, see Farcomeni (2012). We follow the former approach, but leave the distribution for the random effects (coefficients) unspecified. For this purpose, a finite mixture of quantile regression models is introduced, where components of the finite mixture represent clusters of individuals that share homogeneous values of model parameters, conditional on a given quantile. The proposed approach may be considered as a nonparametric alternative to Gaussian or robust random effect models described by Geraci and Bottai (2007), or a particular case of the model introduced by Farcomeni (2012), with a diagonal transition matrix. For its flexibility and ease of implementation, the proposed approach can be easily extended beyond the simple random intercept case, to consider random coefficients with higher dimensions.

There are, however, alternatives to quantile regression, such as M-quantile regression (Breckling and Chambers, 1988), which extend the ideas of M-estimation (Huber, 1964) to a different set of location parameters of the response conditional distribution that lie between quantiles and expectiles. M-quantile regression could be considered as a quantile-like generalisation of mean regression based on influence functions, or a combination of quantile and expectile regression (Newey and Powell, 1987). In fact, M-quantiles aim at combining the robustness properties of quantiles with the efficiency properties of expectiles. **When we turn from quantile to M-quantile models, we *gain* uniqueness of the ML solutions which is, indeed, not granted in the quantile context, and a series of standard tools, mainly based on Fisher information matrices, to check for identifiability and to build standard error estimates for model parameters.** For these reasons, we extend the idea of finite mixture of quantile regression models to M-quantile regression models. **The choice of using an unspecified distribution for the random parameters is based on the empirical evidence that semiparametric models based on a discrete estimate of the mixing distribution are easily handled and estimated, they can be easily linked to clustering purposes, and are robust to departures from Gaussianity.**

We propose to estimate model parameters through Maximum Likelihood (ML); to this aim, we need to calculate the observed and the complete data likelihood function, which, in turn, needs a density function to be specified for the response, conditional on

observed covariates and individual-specific random effects (coefficients). Geraci and Bottai (2007) discuss the relationship between the loss function for quantile regression and the maximisation of a likelihood function formed by combining independently distributed Asymmetric Laplace densities (in the following ALD). A similar alternative for M-quantile regression has been proposed by Bianchi et al. (2015). Although the authors assume that the loss functions belong to the large class of continuously differentiable convex functions, a special attention is devoted to the tilted version of the popular Huber loss function. In particular, Bianchi et al. (2015) propose a new density function called the Asymmetric Least Informative distribution (ALID). They show that the minimisation of the Huber loss function (Huber, 1964) is equivalent to the maximisation of a likelihood function obtained by combining independently distributed Asymmetric Least Informative densities. In this paper, we use the ALID and describe how a finite mixture M-quantile model can be fitted in a ML framework by using the EM algorithm.

To analyse the behaviour of the proposed model parameter estimates on real data, we present the analysis of two benchmark examples. Further, we develop a large scale simulation study where we consider a variety of empirical settings, with varying sample size, cluster size, and distribution of the random effects, in a simple random intercept and a more complex random coefficient context. To give a formal assessment of the sample variability of model parameter estimates, we propose to compute the standard error estimates in the M-quantile setting by the inverse of the observed information matrix calculated using the Oakes (1999) formula; for the quantile case the standard errors are estimated through nonparametric bootstrap. A section of the simulation study is designed to evaluate the behaviour of the standard error estimates in a variety of empirical settings.

The structure of the paper is as follows: in Section 2, we briefly review standard finite mixture models in the context of mixed effect (mean) regression models for responses in the exponential family. In particular, we discuss the approach based on the nonparametric maximum likelihood (NPML) estimation of the mixing distribution. In Section 3, standard M-quantile models for continuous responses are reviewed; given the hierarchical structure in the data, we then propose to extend the finite mixture approach to both quantile and M-quantile regression models. The EM algorithm for ML estimation, the score function and the Hessian corresponding to the proposed model are discussed in Section 4, where the Oakes' formula (Oakes, 1999) is used to derive the standard errors of model parameters. Section 5 entails a large scale simulation study where we analyse the proposed models performance in a variety of scenarios and study the behaviour of the proposed standard error estimators as well. Section 6 contains the analysis of two well-known benchmark datasets to give some further insights on the empirical behaviour of the proposed models. Last section contains concluding remarks and outlines possible future research agenda.

2 Finite Mixtures of regression models

Let y_i, $i = 1, \ldots, n$, be the observed response variable for the i-th subject, and $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})'$ be a p-dimensional column vector of explanatory variables, where, optionally, $x_{i1} \equiv 1$. A generalised linear regression model (GLM) is built by assuming that the observed response is drawn from a density in the exponential family

$$y_i \sim \mathrm{EF}\left(\boldsymbol{\theta}_i\right).$$

The expected value of the response is defined as a function of the observed covariates through a link function:

$$g\left\{E(y_i \mid \boldsymbol{x}_i)\right\} = \boldsymbol{x}_i'\boldsymbol{\beta}. \tag{1}$$

The link function is **an invertible function** such that $g\left(\cdot\right) : M \to \mathbb{R}$, **where M represents the range of the expected value**.

As noted by Aitkin et al. (2005), failure of the model to fit the data could be due to misspecification of any element defining the GLM we have considered so far. However, there is a simple way to unify these possibilities through omitted variables. That is, some fundamental covariates have not been considered in the model specification and, therefore, the regression structure is ill-defined. The omitted covariates could be orthogonal or, more generally, they may be associated with the observed ones. In this last case, their omission may cause a potential bias in the effect estimates for the observed covariates, or cause dependence between observations. To avoid these and related problems, a set of individual-specific random effects $\boldsymbol{b}_i$ is included in the model:

$$g\left\{E(y_i \mid \boldsymbol{x}_i, \boldsymbol{b}_i)\right\} = \boldsymbol{x}_i'\boldsymbol{\beta} + \boldsymbol{w}_i'\boldsymbol{b}_i, \tag{2}$$

where $\boldsymbol{w}_i$ is a subset of the design vector $\boldsymbol{x}_i$ that contains those $p_1 \leqslant p$ variables whose effects are assumed to be individual-specific; that is, the effects $\boldsymbol{b}_i$ vary across individuals according to a distribution $h(\cdot)$. **Usually, it is assumed that $E(\boldsymbol{b}_i) = \boldsymbol{0}$ to ensure identifiability of the corresponding elements of $\boldsymbol{\beta}$. This assumption or, more precisely the assumption that $h(\boldsymbol{b}_i \mid \boldsymbol{x}_i) = h(\boldsymbol{b}_i)$, obviously implies independence of sources of unobserved heterogeneity on the observed covariates.** Such a scheme has been used in a variety of empirical applications; just to mention a few, it can be used to handle overdispersion in single-level data (see e.g. Aitkin, 1996), or to account for dependence in multilevel data (see e.g. Aitkin, 1999; Alfó and Trovato, 2004). Such multilevel data may arise quite naturally as clustered (e.g. pupils nested within classrooms), longitudinal or genuine multivariate data (occasions and/or outcomes nested within individuals). In all these cases, we may consider the observed sample as being composed by lower-level units (pupils, time occasions, outcomes) nested within upper-level units (classrooms, individuals), referred to as clusters.

The notation should be slightly modified to account for this potential hierarchical

structure; let y_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, r_i$ represent the observed response variable for the j-th measurement corresponding to the i-th cluster, and let $\boldsymbol{x}_{ij} = (x_{ij1}, \ldots, x_{ijp})'$ denote the corresponding vector of explanatory variables, where, as before, $x_{ij1} \equiv 1$. In the following, for simplicity, we will mainly refer to the multilevel case; obviously, the cross-sectional case could be obtained when $r_i = 1$, $\forall i = 1, \ldots, n$.

The regression model for multilevel data can be simply described by a slight extension of equation (2):

$$g\left\{E(y_{ij} \mid \boldsymbol{x}_{ij}, \boldsymbol{b}_i)\right\} = \boldsymbol{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{w}'_{ij}\boldsymbol{b}_i. \tag{3}$$

Responses are assumed to be independent conditional on the random vector $\boldsymbol{b}_i$; this is usually referred to as the *local independence* assumption. The corresponding likelihood function has the form:

$$L(\boldsymbol{\Phi}) = \prod_{i=1}^{n} \left\{ \int_{\mathcal{B}} \prod_{j=1}^{r_i} f(y_{ij}|\boldsymbol{x}_{ij}, \boldsymbol{b}_i) \mathrm{d}H(\boldsymbol{b}_i) \right\}, \tag{4}$$

where $\boldsymbol{\Phi}$ denotes the global set of parameters, $H(\cdot)$ is the random coefficient distribution and $\mathcal{B}$ the corresponding support. The terms $\boldsymbol{b}_i$, $i = 1, \ldots, n$, are meant to model unobserved individual-specific (time-invariant) heterogeneity common to each lower-level unit within the same i-th upper-level unit, and, in particular, account for dependence between responses recorded within the same cluster. In the general case, the integral defining the likelihood can not be analytically computed. For ML estimation, we may approximate the integral using ordinary Gaussian Quadrature (GQ), see e.g. Abramowitz and Stegun (1964) and Press et al. (2007); alternatively, we may consider adaptive Gaussian quadrature (aGQ) schemes, based on scaling and centering the $\boldsymbol{b}_i$ domain, see e.g. Liu and Pierce (1994), Pinheiro and Bates (1995). A disadvantage of such approaches lies in the required computational effort, which is exponentially increasing with the dimension of the random parameter vector (both GQ and aGQ), and the need for the calculation of posterior modes and curvatures for each unit in the sample (aGQ only). For these reasons, Monte Carlo and simulated ML approaches have been discussed as potential alternatives, see Geyer and Thompson (1992), McCulloch (1994), Munkin and Trivedi (1999). However, for finite size samples and, in the case of multilevel data, for short individual sequences, these methods may not either provide a good approximation to the *true* mixing distribution (Gueorguieva, 2001), or be inefficient as the missing information increases (Jank and Booth, 2003).

Rather than specifying a parametric distribution for the random effects, we may leave $H(\cdot)$ unspecified and approximate it by using a discrete distribution on $G < n$ locations **$\{\boldsymbol{b}_1, \ldots, \boldsymbol{b}_G\}$, with associated probabilities defined by $\pi_k = \mathrm{Pr}(\boldsymbol{b}_i = \boldsymbol{b}_k)$, $i = 1, \ldots, n$ and $k = 1, \ldots, G$. That is, $\boldsymbol{b}_i \sim \sum_{k=1}^{G} \pi_k \delta_{\boldsymbol{b}_k}$ where δ_θ is a one-point distribution**

putting a unit mass at θ**.** In this case, the likelihood in equation (4) reduces to

$$L(\boldsymbol{\Phi}) = \prod_{i=1}^{n} \left\{ \sum_{k=1}^{G} \prod_{j} f(y_{it}|\boldsymbol{x}_{it}, \boldsymbol{b}_k)\pi_k \right\} =: \prod_{i=1}^{n} \left\{ \sum_{k=1}^{G} \prod_{j} f_{ijk}\pi_k \right\}. \tag{5}$$

As it can be easily noticed, equation (5) resembles the likelihood function for a finite mixture of distributions in the exponential family, where $\boldsymbol{\Phi} = \{\boldsymbol{\beta}, \boldsymbol{b}_1, \dots, \boldsymbol{b}_G, \pi_1, \dots, \pi_G\}$, and f_{ijk} is the distribution of the response variable **for the** j**-th measurement in the** i**-th cluster when** the k-th component of the finite mixture, $k = 1, \dots, G$ is considered. The proposed approach can be derived as a suitable semi-parametric approximation to a fully parametric, possibly continuous, distribution for the random coefficients. It can be also thought of as a model-based clustering approach, where the population of interest is assumed to be divided in G homogeneous sub-populations which differ for the values of the regression parameters, as in Wedel and DeSarbo (1995). The seminal papers by Aitkin (1996, 1999) establish a clear connection between mixed effect models and finite mixtures, by exploiting the theory of Non Parametric Maximum Likelihood (NPML) estimation of a mixing distribution, see Laird (1978).

When compared to a fully parametric framework, the finite mixture approach may be considered as less parsimonious, since the number of unknown parameters to be estimated is higher than in the corresponding parametric model. In fact, locations $\boldsymbol{b}_k$ and masses π_k, $k = 1, \dots, G$, are unknown parameters, as it is G, which is usually treated as fixed and estimated through appropriate penalised likelihood criteria. When using discrete random coefficients, the regression model in equation (3) can be expressed, considering the k-th component of the mixture, as follows:

$$g\left\{E(y_{ij} \mid \boldsymbol{x}_{ij}, \boldsymbol{b}_k)\right\} = \boldsymbol{x}_{ij}'\boldsymbol{\beta} + \boldsymbol{w}_{ij}'\boldsymbol{b}_k. \tag{6}$$

Given the model assumptions, the score function can be written as the posterior expectation of the score function corresponding to a standard GLM:

$$S(\boldsymbol{\Phi}) = \frac{\partial \log[L(\boldsymbol{\Phi})]}{\partial \boldsymbol{\Phi}} = \frac{\partial \ell(\boldsymbol{\Phi})}{\partial \boldsymbol{\Phi}} = \sum_{i=1}^{n} \sum_{k=1}^{G} \left(\frac{f_{ik}\pi_k}{\sum_l f_{il}\pi_l} \right) \sum_{j} \frac{\partial \log f_{ijk}}{\partial \boldsymbol{\Phi}} =: \sum_{i=1}^{n} \sum_{k=1}^{G} \tau_{ik} \sum_{j} \frac{\partial \log f_{ijk}}{\partial \boldsymbol{\Phi}}, \tag{7}$$

where the weights

$$\tau_{ik} = \frac{\prod_{j} f_{ijk}\pi_k}{\sum_l \prod_{j} f_{ijl}\pi_l}$$

represent the posterior probabilities of component membership. Equating (7) to zero gives likelihood equations that are essentially weighted sums of the likelihood equations for a standard GLM, with weights τ_{ik}. While the log-likelihood function can be directly maximised, often ML estimation is based on the use of EM, or EM-type, algorithms. The basic EM algorithm is defined by solving equations for a given set of the weights, and updating the weights as a function of the current parameter estimates.

3   M-quantile regression models

For a continuous (Gaussian) response, the general regression model in (2) is much more flexible than a standard regression model. Nonetheless, it provides a rather incomplete picture, in much the same way as the mean gives an incomplete picture of a distribution. Leaving heterogeneity aside for a moment, quantile regression (Koenker and Bassett, 1978) is defined to summarise the behaviour at different quantiles of the conditional distribution of y given a set of explanatory variables. In the linear case, quantile regression leads to a family of hyper-planes indexed by a real number $q \in (0,1)$ representing the quantile of interest. For example, for $q = 0.05$ the quantile regression hyperplane separates the lowest 5% of the conditional distribution from the remaining 95%. In this sense, quantile regression can be considered as a generalisation of median regression (Koenker and Bassett, 1978), as in the same way expectile regression (Newey and Powell, 1987) is a quantile-like generalisation of standard mean regression. M-quantile regression (Breckling and Chambers, 1988) integrates these concepts within a framework defined by a quantile-like generalisation of regression based on influence functions (M-regression).

The M-quantile of order q for the conditional density of y given the set of covariates $\boldsymbol{x}$, is defined as the solution $MQ_q(y \mid \boldsymbol{x}; \psi)$ of the estimating equation

$$\int \psi_q \left(y - MQ_q(y \mid \boldsymbol{x}; \psi)\right) f(y \mid \boldsymbol{x}) \mathrm{d}y = 0, \tag{8}$$

where $\psi_q(\cdot)$ denotes an asymmetric influence function. In the cross-sectional case, we can say that a linear M-quantile regression model for y_i given the covariate vector $\boldsymbol{x}_i$ is defined by the following expression:

$$MQ_q(y_i \mid \boldsymbol{x}_i; \psi) = \boldsymbol{x}_i' \boldsymbol{\beta}_q. \tag{9}$$

Estimates of $\boldsymbol{\beta}_q$ are obtained by minimizing

$$\sum_{i=1}^{n} \rho_q \left(y_i - MQ_q(y_i \mid \boldsymbol{x}_i; \psi)\right), \tag{10}$$

where $\rho_q(u) = 2\rho(u)\{qI(u > 0) + (1 - q)I(u \le 0)\}$ **and** $d\rho(u)/du = \psi(u)$. **The idea used by Breckling and Chambers (1988) in fitting M-quantile regression is to weight positive residuals by** q **and negative residuals by** $(1 - q)$.

Different types of regression models can be defined according to the choice of the loss function $\rho_q(\cdot)$. In fact, by adopting a particular specification for this function, we may obtain expectile, M-quantile and quantile regression models as particular cases. When $\rho_q(\cdot)$ is the squared loss function, we obtain the standard linear regression model if $q = 0.5$ and the linear expectile regression model if $q \neq 0.5$ (Newey and Powell, 1987). We obtain the linear quantile regression when ρ_q is the asymmetric loss function described by Koenker and Bassett (1978). When $\rho_q(\cdot)$ is the Huber loss function (Huber, 1981), we obtain the M-quantile regression model (Breckling and Chambers, 1988). In the latter case, the function

$\rho_q(\cdot)$ admits first derivative, $\psi_q(\cdot)$, and the minimisation problem can be solved by finding the roots of the following estimating equations

$$\sum_{i=1}^{n} \psi_q(y_i - MQ_q(y_i|\boldsymbol{x}_i; \psi))\boldsymbol{x}_i = \boldsymbol{0}, \tag{11}$$

where

- $\psi_q(\epsilon) = 2\psi(\epsilon/\sigma_q)\{qI(\epsilon > 0) + (1 - q)I(\epsilon \leq 0)\}$;

- $\sigma_q > 0$ is a suitable scale parameter; in the case of robust regression, it is often estimated as $\hat{\sigma}_q = \text{median}|\epsilon|/0.6745$;

- if we use the Huber loss function, $\psi(\epsilon) = \epsilon I(-c \leq \epsilon \leq c) + c \cdot \text{sgn}(\epsilon)I(|\epsilon| > c)$, where c is a tuning constant.

Provided that $c > 0$, estimates of $\boldsymbol{\beta}_q$ are obtained using Iteratively Weighted Least Squares (IWLS). In this case, if we use a continuous influence function (as the Huber function), the IWLS algorithm is known to guarantee convergence to a unique solution, see Kokic et al. (1997). The asymptotic theory for M-quantile regression with i.i.d. errors and fixed regressors can be derived from the results in Huber (1973), see Breckling and Chambers (1988). Bianchi and Salvati (2015) prove the consistency of the estimator of $\boldsymbol{\beta}_q$, provide its asymptotic covariance matrix when regressors are stochastic and propose a variance estimator for the M-quantile regression coefficients based on a sandwich approach.

Quantiles have a more intuitive interpretation than M-quantiles even if they target essentially the same part of the distribution of interest (Jones, 1994). In this paper we focus more on M-estimation because it offers some advantages: (*i*) it easily allows for robust estimation of both fixed and random effects; (*ii*) it can trade robustness for efficiency in inference by selecting the tuning constant of the influence function (quantile and expectile regression can be derived within the M-quantile framework as sketched earlier); (*iii*) it offers computational stability because it can use a wide range of continuous influence functions instead of the absolute value used in quantile regression (Tzavidis et al., 2015).

As discussed earlier (Section 2), the presence of unobserved heterogeneity may lead to poor fit of the model or, in a multilevel data setting, may produce dependence between repeated measurements recorded on the same cluster. **In order to extend M-quantile regression models to account for random coefficients,** the ALID (Asymmetric Least Informative density, Bianchi et al., 2015) is used here as an M-counterpart to ALD (Geraci and Bottai, 2007, 2014; Liu and Bottai, 2009) in standard quantile regression. ALID has the following form:

$$f_q(\cdot) = \frac{1}{B_q(\sigma_q)} \exp\{-\rho_q(\cdot)\} \tag{12}$$

8

where $B_q(\sigma_q)$ is a normalising constant that ensures the density integrates to one:

$$
\begin{aligned}
B_q(\sigma_q) &= \sigma_q\sqrt{\frac{\pi}{q}}\Big[\Phi\Big(c\sqrt{2q}\Big) - 1/2\Big] + \sigma_q\sqrt{\frac{\pi}{1-q}}\Big[\Phi\Big(c\sqrt{2(1-q)}\Big) - 1/2\Big] + \\
&\quad + \frac{\sigma_q}{2cq}\exp\{-c^2 q\} + \frac{\sigma_q}{2c(1-q)}\exp\{-c^2(1-q)\},
\end{aligned}
\tag{13}
$$

where $\Phi(\cdot)$ is the standard Normal CDF. See Bianchi et al. (2015) for details and properties of this density function.

In this paper, we propose to **approximate the distribution of the random co-efficients through a discrete distribution** defined on a finite, G-dimensional, set of locations. According to (6) and conditional on the k-th component of the finite mixture, $k = 1, \ldots, G$, the M-quantile regression model can be written as follows:

$$
MQ_q(y_{ij} \mid \boldsymbol{x}_{ij}, \boldsymbol{b}_{k,q}; \psi) = \boldsymbol{x}_{ij}'\boldsymbol{\beta}_q + \boldsymbol{w}_{ij}'\boldsymbol{b}_{k,q},
\tag{14}
$$

where both fixed and random coefficients vary with $q \in (0,1)$. That is, MQ_q has now to be interpreted conditional on $\boldsymbol{b}_{k,q}$ and the interpretation of model parameters changes accordingly. It is worth noticing that the distribution of the latent variables $\boldsymbol{b}_{k,q}$ may vary with quantiles; for model parameter identifiability, we require that $\sum_k \pi_k \boldsymbol{b}_{k,q} = 0$, as discussed before, while the corresponding covariance matrix varies with q.

To estimate $\boldsymbol{\beta}_q$ and $\boldsymbol{b}_{k,q}$, **we may proceed by differentiating the log-likelihood:**

$$
\ell(\boldsymbol{\Phi}) = \sum_{i=1}^{n}\log\left\{\sum_{k=1}^{G}\prod_j f_q(y_{ij}|\boldsymbol{x}_{ij}, \boldsymbol{b}_k)\pi_k\right\} =: \sum_{i=1}^{n}\log\left\{\sum_{k=1}^{G}f_{ijk}\pi_k\right\},
\tag{15}
$$

where $f_q(\cdot)$ is the ALID in expression (12). Thus, we obtain the following score equations which can be seen to extend the expressions given in (11):

$$
\mathcal{S}\left(\boldsymbol{\beta}_q\right)\sum_{i=1}^{n}\sum_{k=1}^{K}\tau_{ik,q}\sum_{j=1}^{r_i}\psi_q(y_{ij} - \boldsymbol{x}_{ij}'\boldsymbol{\beta}_q - \boldsymbol{w}_{ij}'\boldsymbol{b}_{k,q})\boldsymbol{x}_{ij} = \boldsymbol{0},
\tag{16}
$$

and

$$
\mathcal{S}\left(\boldsymbol{b}_{k,q}\right)\sum_{i=1}^{n}\tau_{ik,q}\sum_{j=1}^{r_i}\psi_q(y_{ij} - \boldsymbol{x}_{ij}'\boldsymbol{\beta}_q - \boldsymbol{w}_{ij}'\boldsymbol{b}_{k,q})\boldsymbol{w}_{ij} = \boldsymbol{0}.
\tag{17}
$$

These likelihood equations represent an extension of the estimating equations in (7), **specified for** the ALID we used to describe the conditional distribution of the response. Similarly, $\tau_{ik,q}$ can be interpreted as the posterior probability that unit i comes from the k-th component of the mixture model at quantile q. It can be computed as

$$
\tau_{ik,q} = \frac{\pi_{k,q}f_{ik,q}}{\sum_l \pi_{l,q}f_{il,q}} = \frac{\pi_{k,q}\prod_j f_{ijk,q}}{\sum_l \pi_{l,q}\prod_j f_{ijl,q}},
\tag{18}
$$

where $f_{ijk,q}$ are computed evaluating the ALID using the parameters corresponding to the k-th component and the q-th quantile. As we have discussed before, we propose to base ML estimation on the EM algorithm; it iterates between two steps, the E-step, where we calculate the posterior probabilities $\tau_{ik,q}$ conditional on observed data and the current parameter estimates, and the M-step where we update parameter estimates with weighted estimation based on current $\tau_{ik,q}$. ML estimation of model parameters is discussed in the following section.

4   Likelihood inference

As it is common in finite mixture models, (indirect) ML estimation is based on the definition of the log-likelihood function for so called *complete* data. For a given quantile $q \in (0, 1)$, we will describe the algorithm in the general case of two-level data, with $j = 1, \ldots, r_i$ measurements and $i = 1, \ldots, n$ upper-level units.

Let $z_{ik,q}$, $i = 1, \ldots, n$, $k = 1, \ldots, G$, denote the indicator variable for the i-th unit in the k-th component of the mixture when the q-th quantile is considered. Each component of the mixture is characterised by a different (sub-) vector of regression coefficients, $\boldsymbol{b}_{k,q}$, $k = 1, \ldots, G$; therefore, we may write $\pi_k = P(z_{ik,q} = 1) = P(\boldsymbol{b}_{i,q} = \boldsymbol{b}_{k,q})$, while the remaining model parameters, $\boldsymbol{\beta}_q$ and σ_q, are assumed to be constant across components.

To be more general, we will discuss random coefficient models, where $\boldsymbol{b}_{k,q}$ represents the effect of covariates in $\boldsymbol{w}_{ij}$; the case of random intercept models can be derived by setting $\boldsymbol{w}_{ij} = 1$. In this context, vectors $\boldsymbol{z}_{i,q} = (z_{i1,q}, \ldots, z_{iG,q})'$, $i = 1, \ldots, n$, are considered as missing data, while the response represents the observed data. Should we have observed, for each unit, the couple $(\boldsymbol{y}_i, \boldsymbol{z}_{i,q})$, the log-likelihood for the *complete* data would have been:

$$
\begin{aligned}
\ell_c(\boldsymbol{\Phi}_q) &= \sum_{i=1}^{n} \sum_{k=1}^{G} z_{ik,q} \left\{ \log \left[f_q(\boldsymbol{y}_i \mid \boldsymbol{\beta}_q, \boldsymbol{b}_{k,q}, \sigma_q) \right] + \log(\pi_{k,q}) \right\} = \\
&= \sum_{i=1}^{n} \sum_{k=1}^{G} z_{ik,q} \left\{ \log \left[\prod_{j=1}^{r_i} f_q(y_{ij} \mid \boldsymbol{\beta}_q, \boldsymbol{b}_{k,q}, \sigma_q) \right] + \log(\pi_{k,q}) \right\} = \\
&= \sum_{i=1}^{n} \sum_{k=1}^{G} z_{ik,q} \left\{ \sum_{j=1}^{r_i} \log \left[f_q(y_{ij} \mid \boldsymbol{\beta}_q, \boldsymbol{b}_{k,q}, \sigma_q) \right] + \log(\pi_{k,q}) \right\}, \quad (19)
\end{aligned}
$$

where $\boldsymbol{\Phi}_q = \left\{ \boldsymbol{\beta}_q, \boldsymbol{b}_{1,q}, \ldots, \boldsymbol{b}_{G,q}, \sigma_q, \pi_{1,q}, \ldots, \pi_{G,q} \right\}$ represents the "global" set of model parameters for the q-th quantile, $q \in (0, 1)$.

Within the E-step of the algorithm, we compute the expected value of the log-likelihood for complete data over the unobservable component indicator vector $\boldsymbol{z}_{i,q}$, conditional on the observed data $\boldsymbol{y}_i$ and the current ML model parameter estimates. That is, the unobservable indicators are replaced by their conditional expectation, which, at the $(t+1)$-

th iteration of the algorithm, are given by

$$\tau_{ik,q}^{(t+1)} = \frac{\pi_{k,q}^{(t)} f_{ik,q}(\widehat{\boldsymbol{\Phi}}_q^{(t)})}{\sum_l \pi_{l,q}^{(t)} f_{il,q}(\widehat{\boldsymbol{\Phi}}_q^{(t)})}, \quad i = 1, \ldots, n, \quad k = 1, \ldots, G. \tag{20}$$

Then, the conditional expectation of the complete data log-likelihood given the observed responses $\boldsymbol{y}_i$ and the current parameter estimates $\widehat{\boldsymbol{\Phi}}_q^{(t)}$ is:

$$Q(\boldsymbol{\Phi}_q \mid \widehat{\boldsymbol{\Phi}}_q^{(t)}) = \mathrm{E}_{\widehat{\boldsymbol{\Phi}}_q^{(t)}} [\ell_c(\boldsymbol{\Phi}_q) \mid \boldsymbol{y}_i] = \sum_{i=1}^n \sum_{k=1}^G \tau_{ik,q}^{(t+1)} \left\{ \sum_{j=1}^{r_i} \log \left[f_q(y_{ij} \mid \boldsymbol{\beta}_q, \boldsymbol{b}_{k,q}, \sigma_q) \right] + \log(\pi_{k,q}) \right\}.$$

By maximising the function $Q(\cdot)$ with respect to $\boldsymbol{\Phi}_q$, we obtain the ML parameter estimates $\widehat{\boldsymbol{\Phi}}_q^{(t+1)}$ given the posterior probabilities $\tau_{ik,q}^{(t+1)}$. For this purpose, in the M-step, the parameter estimates $\boldsymbol{\Phi}_q^{(t+1)}$ are defined to be the solutions to the following score equation:

$$\frac{\partial Q(\boldsymbol{\Phi}_q \mid \widehat{\boldsymbol{\Phi}}_q^{(t)})}{\partial \boldsymbol{\Phi}_q} = \boldsymbol{0},$$

which are equivalent to the score equations for the observed data, $\mathcal{S}(\boldsymbol{\Phi}_q) = \boldsymbol{0}$.

At the $t-$th iteration, we can write the score function by considering each element of $\boldsymbol{\Phi}_q$ separately:

$$
\begin{aligned}
\mathcal{S}(\boldsymbol{\beta}_q) &= \sum_{i=1}^n \sum_{k=1}^G \tau_{ik,q}^{(t)} \sum_{j=1}^{r_i} \frac{\partial}{\partial \boldsymbol{\beta}_q} \log \left[f_q(y_{ij} \mid \boldsymbol{\beta}_q, \boldsymbol{b}_{k,q}, \sigma_q) \right] = \\
&= -\sum_{i=1}^n \sum_{k=1}^G \tau_{ik,q}^{(t)} \sum_{j=1}^{r_i} \frac{\partial \rho_q \left((y_{ij} - \boldsymbol{x}_{ij}' \boldsymbol{\beta}_q - \boldsymbol{w}_{ij}' \boldsymbol{b}_{k,q})/\sigma_q \right)}{\partial \boldsymbol{\beta}_q} = \\
&= \sum_{i=1}^n \sum_{k=1}^G \tau_{ik,q}^{(t)} \sum_{j=1}^{r_i} \psi_q \left(\frac{y_{ij} - \boldsymbol{x}_{ij}' \boldsymbol{\beta}_q - \boldsymbol{w}_{ij}' \boldsymbol{b}_{k,q}}{\sigma_q} \right) \frac{\boldsymbol{x}_{ij}}{\sigma_q} \tag{21}
\end{aligned}
$$

$$
\mathcal{S}(\boldsymbol{b}_{k,q}) = \sum_{i=1}^n \tau_{ik,q}^{(t)} \sum_{j=1}^{r_i} \psi_q \left(\frac{y_{ij} - \boldsymbol{x}_{ij}' \boldsymbol{\beta}_q - \boldsymbol{w}_{ij}' \boldsymbol{b}_{k,q}}{\sigma_q} \right) \frac{\boldsymbol{w}_{ij}}{\sigma_q} \tag{22}
$$

$$
\begin{aligned}
\mathcal{S}(\sigma_q) &= \sum_{i=1}^n \sum_{k=1}^G \tau_{ik,q}^{(t)} \sum_{j=1}^{r_i} \frac{\partial}{\partial \sigma_q} \log \left[f_q(y_{ij} \mid \boldsymbol{\beta}_q, \boldsymbol{b}_{k,q}, \sigma_q) \right] = \\
&= \sum_{i=1}^n \sum_{k=1}^G \tau_{ik,q}^{(t)} \sum_{j=1}^{r_i} \frac{\partial}{\partial \sigma_q} \left\{ - \log(B_q(\sigma_q)) + \log \left[f_q(y_{ij} \mid \boldsymbol{\beta}_q, \boldsymbol{b}_{k,q}, \sigma_q) \right] \right\} = \\
&= \sum_{i=1}^n \sum_{k=1}^G \tau_{ik,q}^{(t)} \sum_{j=1}^{r_i} \left[-\frac{1}{\sigma_q} + \psi_q \left(\frac{y_{ij} - \boldsymbol{x}_{ij}' \boldsymbol{\beta}_q - \boldsymbol{w}_{ij}' \boldsymbol{b}_{k,q}}{\sigma_q} \right) \left(\frac{y_{ij} - \boldsymbol{x}_{ij}' \boldsymbol{\beta}_q - \boldsymbol{w}_{ij}' \boldsymbol{b}_{k,q}}{\sigma_q^2} \right) \right] \tag{23}
\end{aligned}
$$

$$
\mathcal{S}(\pi_k) = \sum_{i=1}^n \left[\frac{\tau_{ik,q}^{(t)}}{\pi_k} - \frac{\tau_{iG,q}^{(t)}}{\pi_G} \right] = \sum_{i=1}^n \left[\frac{\tau_{ik,q}^{(t)}}{\pi_k} - \frac{1 - \sum_{l=1}^{G-1} \tau_{il,q}^{(t)}}{1 - \sum_{l=1}^{G-1} \pi_l} \right], \quad k = 1, \ldots, G-1. \tag{24}
$$

Regardless of the (conditional) density adopted for the response, the expression in equa-

11

tion (24) leads to the following updates for the prior probabilities:

$$\hat{\pi}_{k,q}^{(t)} = \frac{\tau_{ik,q}^{(t)}}{n}.$$

The updated estimates for the remaining model parameters depend on the specific form we choose for the (conditional) density. If we consider the case of finite mixtures of quantile regression models and $f_q(\cdot)$ corresponds to the ALD, a closed form solution for the scale parameter can be derived from expression (23)

$$\hat{\sigma}_q^{(t)} = \frac{1}{\sum_{i=1}^n r_i} \sum_{i=1}^n \sum_{k=1}^G \sum_{j=1}^{r_i} \tau_{ik,q}^{(t)} \rho_q \left(y_{ij} - \boldsymbol{x}_{ij}' \boldsymbol{\beta}_q - \boldsymbol{w}_{ij}' \boldsymbol{b}_{k,q} \right).$$

On the other hand, updated estimates for regression model parameters are obtained by defining an appropriate linear programming algorithm which turns out to be a weighted version of the usual simplex-type algorithm for cross-sectional quantile regression, see e.g. Koenker and D'Orey (1987).

When $f_q(\cdot)$ corresponds to the ALID, the M-step solutions are calculated by an IWLS algorithm, that is essentially a weighted version of the standard algorithm for M-quantile regression models. In this case, the (expected value for the complete data) Hessian is defined by the following expressions (reported only the upper triangular elements):

$$He(\boldsymbol{\beta}_q, \boldsymbol{\beta}_q) = -\sum_{i=1}^n \sum_{k=1}^G \tau_{ik,q}^{(t)} \sum_{j=1}^{r_i} \psi_q' \left(\frac{y_{ij} - \boldsymbol{x}_{ij}' \boldsymbol{\beta}_q - \boldsymbol{w}_{ij}' \boldsymbol{b}_{k,q}}{\sigma_q} \right) \frac{\boldsymbol{x}_{ij}}{\sigma_q} \frac{\boldsymbol{x}_{ij}'}{\sigma_q} \tag{25}$$

$$He(\boldsymbol{\beta}_q, \boldsymbol{b}_{k,q}) = -\sum_{i=1}^n \tau_{ik,q}^{(t)} \sum_{j=1}^{r_i} \psi_q' \left(\frac{y_{ij} - \boldsymbol{x}_{ij}' \boldsymbol{\beta}_q - \boldsymbol{w}_{ij}' \boldsymbol{b}_{k,q}}{\sigma_q} \right) \frac{\boldsymbol{x}_{ij}}{\sigma_q} \frac{\boldsymbol{w}_{ij}'}{\sigma_q} \tag{26}$$

$$He(\boldsymbol{\beta}_q, \sigma_q) = -\sum_{i=1}^n \sum_{k=1}^G \tau_{ik,q}^{(t)} \left[\sum_{j=1}^{r_i} \psi_q' \left(\frac{y_{ij} - \boldsymbol{x}_{ij}' \boldsymbol{\beta}_q - \boldsymbol{w}_{ij}' \boldsymbol{b}_{k,q}}{\sigma_q} \right) \left(\frac{y_{ij} - \boldsymbol{x}_{ij}' \boldsymbol{\beta}_q - \boldsymbol{w}_{ij}' \boldsymbol{b}_{k,q}}{\sigma_q^2} \right) \frac{\boldsymbol{x}_{ij}}{\sigma_q} + \right.$$

$$+ \left. \psi_q \left(\frac{y_{ij} - \boldsymbol{x}_{ij}' \boldsymbol{\beta}_q - \boldsymbol{w}_{ij}' \boldsymbol{b}_{k,q}}{\sigma_q} \right) \frac{\boldsymbol{x}_{ij}}{\sigma_q^2} \right] \tag{27}$$

$$He(\boldsymbol{b}_{k,q}, \boldsymbol{b}_{k,q}) = -\sum_{i=1}^n \tau_{ik,q}^{(t)} \sum_{j=1}^{r_i} \psi_q' \left(\frac{y_{ij} - \boldsymbol{x}_{ij}' \boldsymbol{\beta}_q - \boldsymbol{w}_{ij}' \boldsymbol{b}_{k,q}}{\sigma_q} \right) \frac{\boldsymbol{w}_{ij}}{\sigma_q} \frac{\boldsymbol{w}_{ij}'}{\sigma_q} \tag{28}$$

$$He(\boldsymbol{b}_{k,q}, \sigma_q) = -\sum_{i=1}^n \tau_{ik,q}^{(t)} \left[\sum_{j=1}^{r_i} \psi_q' \left(\frac{y_{ij} - \boldsymbol{x}_{ij}' \boldsymbol{\beta}_q - \boldsymbol{w}_{ij}' \boldsymbol{b}_{k,q}}{\sigma_q} \right) \left(\frac{y_{ij} - \boldsymbol{x}_{ij}' \boldsymbol{\beta}_q - \boldsymbol{w}_{ij}' \boldsymbol{b}_{k,q}}{\sigma_q^2} \right) \frac{\boldsymbol{w}_{ij}}{\sigma_q} + \right.$$

$$+ \left. \psi_q \left(\frac{y_{ij} - \boldsymbol{x}_{ij}' \boldsymbol{\beta}_q - \boldsymbol{w}_{ij}' \boldsymbol{b}_{k,q}}{\sigma_q} \right) \frac{\boldsymbol{w}_{ij}}{\sigma_q^2} \right] \tag{29}$$

$$\tag{30}$$

$$He(\sigma_q, \sigma_q) = \sum_{i=1}^{n} \sum_{k=1}^{G} \tau_{ik,q}^{(t)} \sum_{j=1}^{r_i} \left[\frac{1}{\sigma_q^2} - \psi_q' \left(\frac{y_{ij} - \boldsymbol{x}_{ij}'\boldsymbol{\beta}_q - \boldsymbol{w}_{ij}'\boldsymbol{b}_{k,q}}{\sigma_q} \right) \left(\frac{y_{ij} - \boldsymbol{x}_{ij}'\boldsymbol{\beta}_q - \boldsymbol{w}_{ij}'\boldsymbol{b}_{k,q}}{\sigma_q^2} \right)^2 + \right.$$

$$\left. - 2\psi_q \left(\frac{y_{ij} - \boldsymbol{x}_{ij}'\boldsymbol{\beta}_q - \boldsymbol{w}_{ij}'\boldsymbol{b}_{k,q}}{\sigma_q} \right) \left(\frac{y_{ij} - \boldsymbol{x}_{ij}'\boldsymbol{\beta}_q - \boldsymbol{w}_{ij}'\boldsymbol{b}_{k,q}}{\sigma_q^3} \right) \right] \tag{31}$$

$$He(\pi_k, \pi_l) = \sum_{i=1}^{n} \left[-\frac{\tau_{ik,q}^{(t)}}{\pi_k^2} \mathbf{1}(l=k) - \frac{1 - \sum_{g=1}^{G-1} \tau_{ig,q}^{(t)}}{\left(1 - \sum_{g=1}^{G-1} \pi_g \right)^2} \right] \quad k, l = 1, \ldots, G, \tag{32}$$

where

$$\psi_q'(u) = \begin{cases} 2(1-q) & -c \leq u < 0 \\ 2q & 0 \leq u \leqslant c \\ 0 & |u| > c. \end{cases}$$

The term $\mathbf{1}(A)$ is the indicator function for condition A, and $He(\boldsymbol{\beta}_q, \pi_{k,q}) = He(\boldsymbol{b}_{k,q}, \pi_{k,q}) = He(\sigma_q, \pi_{k,q}) = 0$ due to parameter distinctiveness.

Within the M-step, as it can be easily seen from the expression of the score function, the weights $\tau_{ik,q}^{(t)}$ are kept fixed, and updated at the following E-step. This is essentially the reason why the EM algorithm does not provide estimates of standard errors for model parameters, in that it does not consider the portion of information which is lost due to variability in the $\tau_{ik,q}^{(t)}$. For the quantile regression, this is not a problem, since standard errors are usually computed using non-parametric bootstrap, see Geraci and Bottai (2014). For M-quantile regression, we can rely on the approach discussed by Louis (1982) and, for practical purposes, on the formula described by Oakes (1999). Let us denote by

$$\boldsymbol{I}(\boldsymbol{\Phi}_q) = -\frac{\partial \ell(\boldsymbol{\Phi}_q)}{\partial \boldsymbol{\Phi}_q \partial \boldsymbol{\Phi}_q'}$$

the observed information for $\boldsymbol{\Phi}_q$, and by $\widehat{\boldsymbol{\Phi}}_q^{(t)}$ the current parameter estimates. Then, the Oakes (1999)'s identity states that

$$\boldsymbol{I}(\widehat{\boldsymbol{\Phi}}_q) = - \left\{ \left. \frac{\partial^2 Q(\boldsymbol{\Phi}_q \mid \widehat{\boldsymbol{\Phi}}_q)}{\partial \boldsymbol{\Phi}_q \partial \boldsymbol{\Phi}_q'} \right|_{\widehat{\boldsymbol{\Phi}}_q = \boldsymbol{\Phi}_q} + \left. \frac{\partial^2 Q(\boldsymbol{\Phi}_q \mid \widehat{\boldsymbol{\Phi}}_q)}{\partial \widehat{\boldsymbol{\Phi}}_q \partial \boldsymbol{\Phi}_q'} \right|_{\widehat{\boldsymbol{\Phi}}_q = \boldsymbol{\Phi}_q} \right\}. \tag{33}$$

As it can be noticed, this identity involves two components; the first is given by the conditional expected value of the complete-data Hessian given the observed data, whose elements are reported in expressions (25)–(32) above. This component is simple to obtain from the EM algorithm. The second component involved in expression (33) is the first derivative of the conditional expected value of the complete data score (given the observed data) with respect to the current value of the parameters, that is the posterior weights are not considered as fixed but as functions of the current parameter estimates. For simplicity, we approximate this term by using the R package `numDeriv`.

Once the observed information has been calculated at the ML estimate $\widehat{\boldsymbol{\Phi}}_q$, we proceed

to the calculation of the sandwich estimator (see e.g. Huber, 1967; White, 1980) as follows:

$$\widehat{\text{Cov}}\left(\widehat{\mathbf{\Phi}}_q\right) = \boldsymbol{I}(\widehat{\mathbf{\Phi}}_q)^{-1}\boldsymbol{V}(\widehat{\mathbf{\Phi}}_q)\boldsymbol{I}(\widehat{\mathbf{\Phi}}_q)^{-1},$$

where $\boldsymbol{V}(\widehat{\mathbf{\Phi}}_q) = \sum_{i=1}^n \mathcal{S}_i(\widehat{\mathbf{\Phi}}_q)\mathcal{S}_i(\widehat{\mathbf{\Phi}}_q)'$. Our approach is similar to the procedure described by Friedl and Kauermann (2000), who suggest the use of the sandwich formula to stabilise the observed information estimate by using a Monte Carlo estimate of $\boldsymbol{I}(\widehat{\mathbf{\Phi}}_q)$. This robustified estimate is also protected from the effects of model misspecification that may arise because of the assumption on the distribution of the response variable which is made only to link ML and M-quantile estimation. Using this estimated covariance matrix, we can obtain standard errors for model parameter estimates and check for identifiability in the usual way.

As far as identifiability is of concern, we may observe that standard conditions for identifiability in regression models do not apply directly to finite mixtures of regression models, see DeSarbo and Cron (1988). According to Wang et al. (1996), identifiability needs the covariate matrix to be of full rank. This is rather a necessary condition, since non identifiability may occur also when the full rank condition is fulfilled, see Hennig (2000). According to the latter, the regression parameters are identifiable if and only if the number of components is not higher than the number of distinct $(p-1)$-dimensional hyperplanes in the sample. A similar results is discussed by Follmann and Lambert (1989) for mixed effect logistic models. In this respect, we may notice that model parameters for M-quantile models are estimated through an IWLS-type algorithm; in this sense, standard results for finite mixtures of linear regression models can be extended to finite mixtures of M-quantile regression models as well. However, they can not be applied as they are to finite mixtures of quantile regression models, where singularities may arise also in the homogeneous case (i.e. when $G = 1$), in the presence of categorical covariates. Here, identifiability may be linked to full rank conditions for subsets of the *original* design matrix, and must be checked carefully to avoid aliasing between estimated effects for categorical covariates and estimated component-specific locations.

5   Simulation Study

In this section we describe **an extensive** Monte Carlo simulation study carried out to assess the performance of the proposed finite mixtures of quantile and M-quantile regression models for $q = 0.25$ and $q = 0.50$. Data are generated under the following model

$$y_{ij} = (\beta_1 + b_{1i}) + (\beta_2 + b_{2i})x_{2ij} + \beta_3 x_{3ij} + \varepsilon_{ij}, \quad i = 1, \ldots, n, \ j = 1, \ldots, r_i,$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)' = (100, 2, 1)'$, while b_{1i} and b_{2i} are cluster-specific random parameters. The covariates are generated as

- $x_{2ij} = \delta_i + \zeta_{ij}$, $\delta_i \sim N(0, 1)$, $\zeta_{ij} \sim N(0, 1)$, and

- $x_{3ij} \sim \text{Binom}(1, 0.5)$.

We consider a different number of clusters $n = \{100, 250\}$ and number of measurements for each cluster $r_i = r = \{1, 5\}$. **The sample size varies from** 100 $(r = 1, n = 100)$ **to** **1250** $(r = 5, n = 250)$. The individual-specific random parameters and the error terms are independently generated according to four scenarios:

(NN) $b_{1i} \sim N(0, 1)$, $\varepsilon_{ij} \sim N(0, 1)$;

(TT) $b_{1i} \sim t(3)$, $\varepsilon_{ij} \sim t(3)$;

(NNN) $b_{1i} \sim N(0, 1)$, $b_{2i} \sim N(0, 1)$, $\varepsilon_{ij} \sim N(0, 1)$;

(TTT) $b_{1i} \sim t(3)$, $b_{2i} \sim t(3)$, $\varepsilon_{ij} \sim t(3)$.

Each scenario is replicated $L = 1000$ times. Under scenarios (NN) and (NNN) the assumptions of Gaussian random parameters hold, while for scenarios (TT) and (TTT) these hypotheses are violated, with heavier tails. **In the following, the standard deviation of the random parameters will be denoted by** σ_b **in the univariate case (random intercept) and** σ_{b_l}, $l = 1, 2$ **in the bivariate (random intercept and random slope) case.**

A first aim of this simulation study is to assess the ability of Finite Mixture Quantile Regression (FMQR) and of Finite Mixture M-quantile models (FMMQ) to capture hierarchical effects when modelling the (M-)quantiles of the conditional distribution of the response given the covariates. A second aim is to assess the performance of the standard error estimators for model parameters; as outlined before, the estimates are computed by inverting the observed information obtained using the Oakes (1999) formula. These are compared with the estimates obtained using the sandwich and the non-parametric bootstrap estimators.

As for the first aim, FMQR and FMMQ in (14) are compared with the following models:

- the quantile regression model with Gaussian random effects (QRRE) proposed by Geraci and Bottai (2007), **fitted using the** `lqmm` **function of the** `lqmm` **package of** R;

- the quantile regression model (QR) **fitted by the** `rq` **function of the** `quantreg` **package of** R;

- the M-quantile regression model (MQ) **fitted by a modified version of the** `rlm` **function in** R, **through iteratively re-weighted least squares.**

When $q = 0.5$, we also consider the linear mixed effect model with either a Gaussian (Mixed, see e.g. Pinheiro and Bates, 1995), or a nonparametric specification of the random effect distribution (MixedNP, see e.g. Aitkin, 1999). **Mixed is fitted using the** `lmer`

function from the R package `lmer4`, while MixedNP is fitted using the `allvc` function in the R package `npmlreg`. The EM algorithm for fitting the FMMQ has been implemented in a function written in R and it is available from the authors upon request.

For FMMQ, we use the Huber influence function with $c = 1.345$ (FMMQ$_R$) and its expectile version with $c = 100$ (FMMQ$_E$). **If the data does not present outlying values, the best choice for c is a value that tends to ∞. On the other hand, if the data is heavily contaminated, the value of c should be chosen to be small. Cantoni and Ronchetti (2001) suggest using $c = 1.2$ and Street et al. (1988) use $c = 1.25$. Huber (1981) suggests $c = 1.345$ because it gives reasonably high efficiency in the normal case; it produces 95% efficiency when the errors are normal and still offers protection against outliers. Wang et al. (2007) propose a data-driven approach to automatically adjust the value of the tuning constant to provide the necessary resistance against outliers. For a given family of loss functions, such as the Huber family, Wang et al. (2007) shows that the best value for c can be obtained from the data so that the asymptotic efficiency is maximized. For FMMQ, the choice of the tuning constant could be done automatically using the likelihood function in the M-step of the EM algorithm. In this step $\boldsymbol{\beta}_q$, $\mathbf{b}_{k,q}$, σ_q and c can be estimated simultaneously by maximizing the log-likelihood function. Unfortunately there is no closed form solution for the tuning constant c. For this reason, a data-driven procedure to estimate c by means of the log-likelihood function can be carried out by defining a grid of c values $-$ e.g., $g \in (0.01, 100)$ $-$ computing the profile log-likelihood for a fixed $c \in g$ conditional on $\boldsymbol{\beta}_q$, $\mathbf{b}_{k,q}$, σ_q. Then c can be selected as the value from the grid that maximises the profile log-likelihood function. This procedure can be computationally intensive and, for this reason, we have decided to select only two different values, $c = 1.345$ ad $c = 100$.**

FMMQ$_R$ ($c = 1.345$) is robust to outliers and, when clustering is present, we may guess that it should perform better than the single level M-quantile model, MQ. Also, when outliers are present, we expect that FMMQ$_R$ will perform better than both its expectile counterpart, FMMQ$_E$ ($c = 100$), and the linear random effect model (Mixed).

For each simulated sample, we employ a deterministic strategy for initialisation, based on considering masses and locations of standard Gauss-Hermite quadrature as starting points. Obviously, this strategy can be substantially improved by adopting a multi-start random initialisation, as the one we have used in the analysis of real data examples (see Section 6). However, this strategy may significantly increase the global computational burden and, for this reason, it is not employed in this large scale simulation study. **For each simulated sample, the number of latent states for FMMQ$_E$, FMMQ$_R$ and FMQR is selected over the interval $G = \{2, ..., 6\}$ through the AIC criterion.**

To compare the different models, we mainly focus on the estimation of the fixed effect

parameters; for each fixed effect parameter, the performance is assessed using the following indicators.

(a) Average Relative Bias (ARB) defined, for a generic regression coefficient estimator $\hat{\beta}$, by

$$ARB(\hat{\beta}) = L^{-1} \sum_{l=1}^{L} \frac{\hat{\beta}^{(l)} - \beta}{\beta} \times 100,$$

where $\hat{\beta}^{(l)}$ is the estimate from the l-th sample and β is the corresponding true value. For quantile and M-quantile regression, the true value is obtained from the corresponding quantile of the conditional distribution of the response given the covariates.

(b) Efficiency (EFF) with respect to FMMQ_R, defined by

$$EFF(\hat{\beta}) = \frac{S^2(\hat{\beta})}{S^2_{\text{FMMQ}_R}(\hat{\beta})},$$

where $S^2(\hat{\beta}) = L^{-1} \sum_{l=1}^{L} (\hat{\beta}^{(l)} - \bar{\beta})^2$, $\bar{\beta} = L^{-1} \sum_{l=1}^{L} \hat{\beta}^{(l)}$. We have used FMMQ_R as a reference because we are mainly interested in testing its ability to account for the hierarchical structure in the data.

For each simulated scenario, we report the point estimates for $\boldsymbol{\beta}_q$, $\sigma_{b_1}^2$, $\sigma_{b_2}^2$ and for the scale σ_q, all averaged over the simulated samples.

We start by discussing the scenarios under a simple random intercept model, i.e. cases (NN) and (TT) described above, where $\sigma_{b_2}^2 = 0$. Tables 1 and 2 show that FMMQ_E and FMMQ_R generally perform better than the competitors across the two simulated scenarios. In particular, under the (NN) scenario (see Table 1), large gains in efficiency of the FMMQ_E and FMMQ_R over all the competing models can be noticed for $n = 250$ and $r = 5$, as expected. At $q = 0.5$, Mixed shows the best performance in terms of efficiency for all combinations of n and r. Nonetheless, for increasing sample size and number of measurements per cluster, the proposed models are only slightly worse than Mixed, even if the latter represents the *true* data generating process. At $q = 0.25$ and $r = 5$, FMMQ_E is also slightly more efficient than FMMQ_R, and this finding proves the ability of the FMMQ_E in extending the linear mixed model to further quantiles of the conditional distribution.

The bias of the estimate of the intercept term is particularly high when $q = 0.25$ for all models but FMQR. FMMQ_E and FMMQ_R exhibit a good performance when we consider the estimation of $\sigma_{b_1}^2$ and σ_q, especially as n and r increase. The superior performance in terms of bias and efficiency of the robust models can be clearly observed when we consider heavy tailed distribution for the random effects and the error terms, under the (TT) scenario. By looking at Table 2, we can notice large gains in efficiency for the robust models, FMQR and FMMQ_R, in particular for $r = 5$; as expected, MQ performs well in terms of efficiency when $r = 1$. On the contrary, QRRE shows large bias and variability

that could be due to some instability in the estimation algorithm. These results illustrate that the aims of the robust models (FMMQ$_R$ and FMQR) are achieved; that is, these models protect us against outlying values when modelling the quantiles of the conditional distribution of y given the covariates and account for the hierarchical structure in the data.

Tables 3 and 4 report the results under the (NNN) and the (TTT) scenarios with $n = 100$ and $r = (1, 5)$. Results for $n = 250$ are not reported for reasons of space but are available from the authors. In scenario (NNN), FMMQ$_R$ and QRRE perform well in terms of efficiency and bias when $r = 1$ at $q = 0.5$; as expected, as the number of within-cluster measurements increases (i.e. $r = 5$) FMMQ$_E$ shows the best performance. At $q = 0.5$ and $r = 1$ QRRE, Mixed and MQ have the best performance in terms of bias and efficiency. The good behaviour of MQ can be explained by the fact that it uses a more parsimonious model that does not take into account the hierarchical structure and, therefore, the final estimates are less variable than those obtained by FMMQ$_R$. When r increases, Mixed, MixedNP and FMMQ$_E$ show the best performance both in terms of bias and efficiency. As r increases, all models that account for the hierarchical structure of the data improve with respect to the estimation of the variance components and the scale, but MixedNP, which seems to have some problems in estimating $\sigma_{b_1}^2$. Coherently with the findings we have derived from Table 2, the superior performance of robust models, especially FMMQ$_R$ and FMQR, is also replicated under the (TTT) scenario (Table 4).

The number of components G chosen for FMMQ$_E$, FMMQ$_R$, **FMQR over the interval $G = \{2, ..., 6\}$ according to the lowest AIC value is 2 in about 60% of runs when $r = 1$, whereas it becomes 4 or 5 in about 80% of cases when $r = 5$ ($q = 0.25, 0.50$) and $n = 100, 250$ under the Gaussian scenarios (NN and NNN). Under (TT) and (TTT) scenarios for $q = 0.25, 0.50$, the number of components selected through AIC is 2 in about 60% of runs when the robust models (**FMMQ$_R$, **FMQR) are fitted and $r = 1$, whereas the number of components is 2 or 3 in about 70% of cases when** FMMQ$_E$ **is fitted. When $r = 5$ the modal value for robust models is $G = 5$ while it is $G = 4$ for** FMMQ$_E$.

To evaluate the speed of convergence for the EM algorithm presented in Section 3, we report in Table 7 the average time to convergence for FMMQ$_E$, FMMQ$_R$ **and FMQR models with ($G = 2, r = 1$) and ($G = 4, r = 5$) under Gaussian (NN) and t_3 (TT) data generating scenarios. The time to convergence refers to an optimised** R **code on a laptop with a 2 GHz Intel Core i7 and 16 Gb RAM. Table 7 shows also the convergence time of the fitted algorithm for QRRE, MQ, QR, Mixed and MixedNP models. The** FMMQ$_R$ **model shows good computational performance when the number of clusters and the number of measurements for each cluster increase. The average time to convergence obviously increases with G and the sample size, but it looks reasonable in all cases except for** FMMQ$_E$ **when $n = 250$ and $r = 5$.**

Scenarios (NN) and (TT) have also been used to evaluate the standard error estimates derived according to the procedure illustrated in Section 4, either when the assumption of Gaussian random effects holds or when this assumption is violated. To start with, Tables 5 and 6 provide results on how well the estimates for the standard errors of the fixed effect β_1 and for the scale σ_q approximate the corresponding 'true' (Monte Carlo) variance. The results are based on $n = 100$ and $B = 1000$ Monte Carlo replications; for sake of brevity, the results for FMMQ$_E$ are reported under the (NN) scenario while the results for FMMQ$_R$ are reported under the (TT) scenario. The bootstrap and the sandwich estimates for standard error of the regression coefficient provide a good approximation to the true (Monte Carlo) variance, whereas the analytic estimates for FMMQ$_R$ and FMMQ$_E$ show a slight underestimation when $r = 1$ under both scenarios, while the performance is quite reliable when $r = 5$. The analytical, the sandwich and the bootstrap variance estimates for the scale parameter (Table 6) show a large underestimation of the empirical variance when $r = 1$, due to a clear multimodality of the corresponding Monte Carlo distribution which can be ascribed to a lack of a proper initialisation strategy. The performance of the variance estimates improves as r increases. Considering the results, we have decided to use the sandwich and the bootstrap estimators to compute the standard errors for the applications in Section 6.

6 Case Studies

In this section, we present the application of the Finite Mixture of Quantile (FMQR) and M-quantile (FMMQ$_R$ and FMMQ$_E$) regression models to data from two benchmark case studies. In the first case study, we analyse repeated measures data on labour pain. These data have been discussed by Davis (1991) and subsequently analysed by Jung (1996), Geraci and Bottai (2007) and Farcomeni (2012). In the second case study, we analyse a placebo-controlled, double-blind, randomised trial conducted by the Treatment of Lead-Exposed Children (TLC) Trial Group (2000).

6.1 Case Study 1: Analysis of Pain Labor data

This data set consists of repeated measures of self-reported amount of pain for 83 women in labor, 43 of them randomly assigned to a pain medication group and 40 to a placebo group. The outcome variable is the amount of pain measured every 30 min on a 100-mm line, where 0 means no pain and 100 means extreme pain. The maximum number of measurements (r_i) for a woman is six and there are, in total, $\sum_{i=1}^{n} r_i = 357$ observations. The explanatory variables are the `Treatment` indicator (treatment=1, placebo=0) and the measurement occasion `Time`, with the category (`Time` $= 0$) being been chosen as reference. These data are severely skewed, and the skewness changes magnitude, and even sign, over time. Figures 1 and 2 show selected diagnostics for a linear mixed model fit to the response variable based on the original Pain Labor data using `Treatment`, `Time`

Table 1: Random intercepts, Gaussian data generating scenario (NN). Values of bias (ARB), efficiency (EFF), and Monte Carlo average of point estimates for $\sigma^2_{b_1}$ and σ_q at $q = 0.25$ and $q = 0.5$. The results are based on 1000 Monte Carlo replications.

| | $r = 1$ | | | | | | $r = 5$ | | | | | |
| | $\hat{\beta}_1$ | | $\hat{\beta}_2$ | | | | $\hat{\beta}_1$ | | $\hat{\beta}_2$ | | | |
Model	ARB	EFF	ARB	EFF	$\bar{\hat{\sigma}}^2_{b_1}$	$\bar{\hat{\sigma}}_q$	ARB	EFF	ARB	EFF	$\bar{\hat{\sigma}}^2_{b_1}$	$\bar{\hat{\sigma}}_q$
$q = 0.25$						$n = 100$						
QRRE	0.57	0.84	-0.04	1.04	0.98	1.00	0.44	1.68	-0.00	1.72	1.00	1.27
FMMQ$_E$	0.27	1.13	0.00	1.11	0.97	0.95	0.23	0.98	0.00	0.99	0.99	1.04
FMMQ$_R$	0.25	1.00	0.00	1.00	0.99	1.17	0.24	1.00	0.00	1.00	1.00	1.28
FMQR	-0.00	1.18	0.00	1.51	0.99	1.08	-0.00	1.21	0.00	1.51	1.01	1.27
MQ	-0.00	1.32	0.00	1.86	-	-	0.01	1.41	0.00	1.85	-	-
QR	0.00	1.73	-0.00	2.72	-	-	-0.28	2.07	0.00	2.95	-	-
						$n = 250$						
QRRE	0.37	1.21	-0.02	1.44	0.98	1.04	0.24	2.63	-0.01	1.74	0.99	1.27
FMMQ$_E$	0.23	1.23	0.00	1.09	0.88	1.06	0.24	0.99	-0.00	0.98	1.00	1.03
FMMQ$_R$	0.32	1.00	0.00	1.00	1.09	1.05	0.25	1.00	-0.0	1.00	1.00	1.27
FMQR	0.11	1.62	0.00	1.69	1.09	0.96	-0.00	1.28	-0.00	1.59	1.02	1.27
MQ	0.01	1.01	0.00	1.14	-	-	0.02	1.30	-0.00	1.66	-	-
QR	-0.27	1.59	0.00	1.73	-	-	-0.27	1.91	0.00	2.56	-	-
$q = 0.50$						$n = 100$						
Mixed	-0.01	0.80	0.00	0.75	0.00	1.41	-0.00	0.97	0.00	0.96	1.00	1.00
MixedNP	-0.02	1.37	0.00	1.44	1.34	0.41	-0.00	1.00	0.00	1.00	0.99	1.00
QRRE	-0.01	0.80	0.00	0.79	0.97	1.00	-0.00	1.24	0.00	1.68	0.99	1.09
FMMQ$_E$	-0.01	1.02	0.00	1.03	0.97	0.91	-0.00	1.00	0.00	1.02	0.99	1.00
FMMQ$_R$	-0.02	1.00	0.00	1.00	1.02	0.90	-0.00	1.00	0.00	1.00	0.98	1.04
FMQR	-0.02	1.13	0.00	1.14	0.96	0.95	-0.00	1.18	0.00	1.51	0.99	1.08
MQ	-0.01	0.84	0.00	0.79	-	-	-0.00	1.32	0.00	1.86	-	-
QR	0.00	1.26	0.00	1.18	-	-	0.00	1.73	-0.00	2.72	-	-
						$n = 250$						
Mixed	0.00	0.89	0.00	0.86	0.00	1.41	0.00	0.98	0.00	0.96	1.00	1.00
MixedNP	0.00	1.40	0.00	1.51	1.31	0.49	0.00	0.99	0.00	0.98	1.00	1.00
QRRE	0.02	0.90	0.00	0.93	0.93	1.06	0.00	1.20	-0.00	1.35	0.98	1.09
FMMQ$_E$	0.00	1.00	0.00	1.00	0.83	1.06	0.00	0.99	0.00	0.98	1.00	1.00
FMMQ$_R$	0.00	1.00	0.00	1.00	0.98	0.98	0.00	1.00	0.00	1.00	0.99	1.04
FMQR	0.00	1.28	0.00	1.35	1.00	0.96	0.00	1.19	-0.00	1.37	1.00	1.09
MQ	-0.00	0.96	0.00	0.90	-	-	0.00	1.22	0.00	1.54	-	-
QR	-0.00	1.50	0.00	1.40	-	-	0.00	1.65	0.00	2.22	-	-

Table 2: Random intercepts, t_3 data generating scenario (TT). Values of bias (ARB), efficiency (EFF), and Monte Carlo average of point estimates for $\sigma^2_{b_1}$ and σ_q at $q = 0.25$ and $q = 0.5$. The results are based on 1000 Monte Carlo replications.

| | $r = 1$ | | | | | | $r = 5$ | | | | | |
| | $\hat{\beta}_1$ | | $\hat{\beta}_2$ | | | | $\hat{\beta}_1$ | | $\hat{\beta}_2$ | | | |
Model	ARB	EFF	ARB	EFF	$\bar{\hat{\sigma}}^2_{b_1}$	$\bar{\hat{\sigma}}_q$	ARB	EFF	ARB	EFF	$\bar{\hat{\sigma}}^2_{b_1}$	$\bar{\hat{\sigma}}_q$
$q = 0.25$							$n = 100$					
QRRE	0.46	1.06	-0.06	1.09	1.01	2.25	-0.19	1.74	-0.00	0.98	1.55	2.37
FMMQ_E	0.17	1.11	-0.00	0.98	1.71	1.38	0.00	1.08	0.00	1.32	1.61	1.55
FMMQ_R	0.03	1.00	-0.00	1.00	1.31	1.88	0.00	1.00	0.00	1.00	1.61	1.92
FMQR	-0.29	1.11	-0.00	1.16	1.04	2.14	0.00	0.91	0.00	0.91	1.48	1.61
MQ	-0.13	0.98	-0.00	0.99	-	-	0.00	0.83	0.00	1.51	-	-
QR	-0.44	1.24	-0.00	1.28	-	-	0.00	0.91	0.00	1.89	-	-
							$n = 250$					
QRRE	0.05	1.16	-0.03	1.29	1.05	2.22	0.09	1.43	-0.00	1.09	1.47	1.93
FMMQ_E	0.12	1.11	-0.00	0.99	1.73	1.44	0.10	1.17	0.00	1.20	1.73	1.49
FMMQ_R	0.05	1.00	-0.00	1.00	1.44	1.83	0.13	1.00	0.00	1.00	1.69	1.78
FMQR	-0.32	1.16	0.00	1.33	1.20	2.09	-0.15	0.91	0.00	0.96	1.54	1.96
MQ	-0.13	0.96	-0.00	1.07	-	-	-0.15	1.00	0.00	1.37	-	-
QR	-0.43	1.26	-0.00	1.40	-	-	-0.44	1.16	0.00	1.77	-	-
$q = 0.50$							$n = 100$					
Mixed	0.01	1.40	-0.00	1.36	0.00	2.36	0.02	1.07	-0.00	1.12	1.64	1.68
MixedNP	0.02	1.41	-0.00	1.50	2.19	0.81	0.02	1.08	-0.00	1.18	1.64	1.67
QRRE	0.02	0.95	-0.00	0.93	0.61	2.09	0.00	1.67	-0.00	0.86	1.51	1.58
FMMQ_E	0.02	1.15	-0.00	1.13	1.75	1.43	0.02	1.09	-0.00	1.19	1.62	1.63
FMMQ_R	0.03	1.00	-0.00	1.00	0.90	1.73	0.01	1.00	-0.00	1.00	1.62	1.51
FMQR	0.01	0.97	-0.00	0.99	0.56	2.11	0.01	0.84	-0.00	0.90	1.44	1.62
MQ	0.02	0.85	-0.00	0.85	-	-	0.00	0.83	-0.00	1.27	-	-
QR	0.03	1.07	-0.00	1.05	-	-	0.00	0.95	0.00	1.62	-	-
							$n = 250$					
Mixed	0.01	1.36	-0.00	1.48	0.00	2.41	0.00	1.08	0.00	1.22	1.66	1.71
MixedNP	0.02	1.17	-0.00	1.16	2.07	1.19	0.00	1.06	0.00	1.21	1.65	1.72
QRRE	0.02	0.98	-0.00	1.04	0.63	2.13	-0.00	1.92	0.00	0.94	1.50	1.59
FMMQ_E	0.01	1.07	-0.00	1.10	1.71	1.59	0.00	1.06	0.00	1.22	1.65	1.64
FMMQ_R	0.02	1.00	-0.00	1.00	0.65	1.95	0.00	1.00	0.00	1.00	1.62	1.52
FMQR	0.02	1.03	-0.00	1.10	0.56	2.16	0.00	0.90	0.00	0.90	1.45	1.62
MQ	0.02	0.90	-0.00	0.95	-	-	0.00	0.90	0.00	1.29	-	-
QR	0.01	1.12	-0.00	1.18	-	-	-0.00	0.96	0.00	1.65	-	-

Table 3: Random intercepts and slopes, Gaussian data generating scenario (NNN). Values of bias (ARB), efficiency (EFF), and Monte Carlo average of point estimates for $\sigma_{b_1}^2$, $\sigma_{b_2}^2$ and σ_q at $q = 0.25$ and $q = 0.5$. The results are based on 1000 Monte Carlo replications.

Model	$\hat{\beta}_1$ ARB	$\hat{\beta}_1$ EFF	$\hat{\beta}_2$ ARB	$\hat{\beta}_2$ EFF	$\hat{\beta}_3$ ARB	$\hat{\beta}_3$ EFF	$\bar{\hat{\sigma}}_{b_1}^2$	$\bar{\hat{\sigma}}_{b_2}^2$	$\bar{\hat{\sigma}}_q$
$q = 0.25$				$n = 100$, $r = 1$					
QRRE	-0.21	1.00	0.00	1.18	-0.00	0.93	0.51	0.89	1.34
FMMQ$_E$	0.33	1.20	0.01	1.07	-0.05	1.31	1.21	0.71	0.86
FMMQ$_R$	0.19	1.00	0.00	1.00	-0.03	1.00	0.98	0.59	1.43
FMQR	-0.11	1.40	0.01	1.08	-0.03	1.41	1.09	0.51	1.33
MQ	-0.19	0.92	0.00	1.18	-0.02	1.02	-	-	-
QR	-0.53	1.27	0.01	1.52	-0.02	1.53	-	-	-
				$n = 100, r = 5$					
QRRE	-0.07	1.19	0.01	1.90	0.02	0.98	0.99	0.96	1.22
FMMQ$_E$	0.20	0.89	0.01	0.95	0.01	0.87	0.99	0.90	1.13
FMMQ$_R$	0.20	1.00	0.01	1.00	0.01	1.00	1.03	0.82	1.46
FMQR	-0.11	1.19	0.00	1.12	0.01	1.48	1.10	0.74	1.47
MQ	-0.20	1.36	0.00	1.61	0.01	2.03	-	-	-
QR	-0.55	1.68	0.00	1.85	0.01	2.81	-	-	-
$q = 0.5$				$n = 100, r = 1$					
Mixed	0.01	0.92	0.00	0.96	-0.02	0.75	0.00	0.00	1.98
MixedNP	0.01	1.23	-0.05	1.31	-0.01	1.30	1.59	0.72	0.58
QRRE	0.01	0.78	0.01	0.97	-0.02	0.67	0.40	0.61	1.27
FMMQ$_E$	0.02	1.04	0.01	1.01	-0.04	1.08	1.15	0.73	0.87
FMMQ$_R$	0.03	1.00	0.01	1.00	-0.05	1.00	1.22	0.77	0.83
FMQR	0.02	0.98	0.01	1.00	-0.04	0.94	0.94	0.57	1.26
MQ	0.01	0.81	0.00	0.93	-0.02	0.69	-	-	-
QR	0.02	1.11	0.01	1.21	-0.03	0.92	-	-	-
				$n = 100, r = 5$					
Mixed	-0.00	0.81	0.01	0.91	0.01	0.75	0.99	1.00	1.00
MixedNP	-0.00	0.93	0.01	1.00	0.01	0.95	0.94	0.93	1.10
QRRE	-0.00	1.18	0.01	1.58	0.01	0.94	0.96	0.95	1.07
FMMQ$_E$	-0.00	0.94	0.00	1.00	0.01	0.95	0.97	0.91	1.10
FMMQ$_R$	0.00	1.00	0.01	1.00	0.01	1.00	0.97	0.90	1.15
FMQR	-0.00	1.13	0.01	1.09	0.00	1.53	1.02	0.79	1.24
MQ	-0.00	1.39	0.00	1.52	0.01	2.01	-	-	-
QR	-0.00	1.64	0.01	1.69	0.01	2.97	-	-	-

Table 4: Random intercepts and slopes, t_3 data generating scenario (TTT). Values of bias (ARB), efficiency (EFF), and Monte Carlo average of point estimates for $\sigma^2_{b_1}$, $\sigma^2_{b_2}$ and σ_q at $q = 0.25$ and $q = 0.5$. The results are based on 1000 Monte Carlo replications.

Model	$\hat{\bar{\beta}}_1$ ARB	$\hat{\bar{\beta}}_1$ EFF	$\hat{\bar{\beta}}_2$ ARB	$\hat{\bar{\beta}}_2$ EFF	$\hat{\bar{\beta}}_3$ ARB	$\hat{\bar{\beta}}_3$ EFF	$\bar{\hat{\sigma}}^2_{b_1}$	$\bar{\hat{\sigma}}^2_{b_2}$	$\bar{\hat{\sigma}}_q$
$q = 0.25$				$n = 100, r = 1$					
QRRE	-0.66	0.87	0.01	1.08	0.01	0.94	0.48	1.17	2.55
FMMQ$_\mathrm{E}$	0.14	1.17	0.01	1.06	0.00	1.31	2.30	1.28	1.16
FMMQ$_\mathrm{R}$	-0.14	1.00	0.01	1.00	0.00	1.00	1.56	1.08	1.88
FMQR	-0.71	1.02	0.01	0.92	0.01	1.17	1.18	0.74	2.66
MQ	-0.56	0.88	0.01	0.90	-0.01	1.10	-	-	-
QR	-0.92	1.06	0.01	1.09	-0.01	1.35	-	-	-
				$n = 100, r = 5$					
QRRE	-0.38	1.33	0.00	1.88	0.02	0.67	1.56	1.55	1.88
FMMQ$_\mathrm{E}$	-0.10	1.04	0.01	0.99	0.01	1.05	1.72	1.39	1.54
FMMQ$_\mathrm{R}$	-0.08	1.00	0.01	1.00	0.01	1.00	1.84	1.18	1.97
FMQR	-0.48	1.05	0.01	0.98	0.00	1.04	1.72	1.03	2.33
MQ	-0.57	1.01	0.01	1.04	0.01	1.63	-	-	-
QR	-0.93	1.14	0.00	1.08	0.00	2.16	-	-	-
$q = 0.5$				$n = 100, r = 1$					
Mixed	0.01	1.20	0.00	1.24	-0.03	1.32	0.00	0.00	3.37
MixedNP	-0.00	1.54	-0.16	1.37	-0.01	1.25	3.23	1.05	0.95
QRRE	0.01	0.69	0.01	0.82	-0.01	0.65	0.22	0.91	2.52
FMMQ$_\mathrm{E}$	0.01	1.05	0.00	1.07	0.00	1.18	2.29	1.26	1.24
FMMQ$_\mathrm{R}$	0.01	1.00	0.01	1.00	-0.00	1.00	1.73	1.34	1.51
FMQR	0.01	0.76	0.02	0.74	-0.01	0.75	0.59	0.64	2.64
MQ	0.01	0.68	0.01	0.69	-0.01	0.68	-	-	-
QR	0.00	0.80	0.02	0.87	-0.01	0.77	-	-	-
				$n = 100, r = 5$					
Mixed	-0.00	0.86	0.01	0.98	0.00	0.91	1.65	1.64	1.69
MixedNP	0.00	1.04	0.00	1.03	0.00	1.10	1.65	1.44	1.91
QRRE	0.00	1.04	0.02	1.92	0.01	0.69	1.50	1.52	1.61
FMMQ$_\mathrm{E}$	-0.00	1.04	0.01	1.04	0.00	1.15	1.67	1.40	1.88
FMMQ$_\mathrm{R}$	-0.01	1.00	0.01	1.00	0.01	1.00	1.69	1.31	1.81
FMQR	0.00	0.93	0.01	0.91	0.00	0.94	1.53	1.07	1.96
MQ	-0.01	0.87	0.00	0.88	0.02	1.50	-	-	-
QR	-0.01	0.89	0.00	0.99	0.02	1.74	-	-	-

Table 5: Gaussian and t_3 data generating process. Empirical, $S^2(\cdot)$, analytical, $S^2_A(\cdot)$, sandwich, $S^2_S(\cdot)$, and bootstrap, $S^2_{boot}(\cdot)$, variance estimates for the regression parameter estimates at $q = 0.25$ and $q = 0.5$. The results are based on $n = 100$ and $L = 1000$ Monte Carlo replications.

Model	$r=1$				$r=5$			
	$S^2(\hat{\beta}_2)$	$\hat{S}^2_A(\hat{\beta}_2)$	$\hat{S}^2_S(\hat{\beta}_2)$	$\hat{S}^2_{Boot}(\hat{\beta}_2)$	$S^2(\hat{\beta}_2)$	$\hat{S}^2_A(\hat{\beta}_2)$	$\hat{S}^2_S(\hat{\beta}_2)$	$\hat{S}^2_{Boot}(\hat{\beta}_2)$
$q = 0.25$				Gaussian distribution				
FMMQ$_E$	0.0291	0.0209	0.0246	0.0291	0.0090	0.0085	0.0087	0.0091
FMQR	0.0310	-	-	0.0304	0.0105	-	-	0.0108
$q = 0.5$								
FMMQ$_E$	0.0294	0.0211	0.0230	0.0281	0.0088	0.0085	0.0084	0.0086
FMQR	0.0297	-	-	0.0304	0.0112	-	-	0.0109
$q = 0.25$				t_3 distribution				
FMMQ$_R$	0.0391	0.0370	0.0431	0.0414	0.0144	0.0122	0.0139	0.0145
FMQR	0.0433	-	-	0.0479	0.0150	-	-	0.0164
$q = 0.5$								
FMMQ$_R$	0.0369	0.0333	0.0383	0.0381	0.0148	0.0123	0.0124	0.0132
FMQR	0.0358	-	-	0.0404	0.0151	-	-	0.0164

Table 6: Gaussian and t_3 data generating process. Empirical, $S^2(\cdot)$, analytical, $S^2_A(\cdot)$, sandwich, $S^2_S(\cdot)$, and bootstrap, $S^2_{boot}(\cdot)$, variance estimates for the scale parameter estimates at $q = 0.25$ and $q = 0.5$. The results are based on $n = 100$ and $L = 1000$ Monte Carlo replications.

Model	$r=1$				$r=5$			
	$S^2(\hat{\sigma}_q)$	$\hat{S}^2_A(\hat{\sigma}_q)$	$\hat{S}^2_S(\hat{\sigma}_q)$	$\hat{S}^2_{Boot}(\hat{\sigma}_q)$	$S^2(\hat{\sigma}_q)$	$\hat{S}^2_A(\hat{\sigma}_q)$	$\hat{S}^2_S(\hat{\sigma}_q)$	$\hat{S}^2_{Boot}(\hat{\sigma}_q)$
$q = 0.25$				Gaussian distribution				
FMMQ$_E$	0.2818	0.0970	0.1182	0.1064	0.0311	0.0311	0.0316	0.0317
FMQR	0.3131	-	-	0.1659	0.0416	-	-	0.0404
$q = 0.5$								
FMMQ$_E$	0.3032	0.1259	0.1382	0.1150	0.0361	0.0343	0.0345	0.0343
FMQR	0.3108	-	-	0.2661	0.0450	-	-	0.0406
$q = 0.25$				t_3 distribution				
FMMQ$_R$	0.2361	0.1998	0.1700	0.1952	0.0706	0.0496	0.0757	0.0643
FMQR	0.5194	-	-	0.4637	0.1607	-	-	0.1480
$q = 0.5$								
FMMQ$_R$	0.4923	0.1593	0.2312	0.2719	0.0716	0.0532	0.0718	0.0694
FMQR	0.3599	-	-	0.3162	0.1607	-	-	0.1480

Table 7: Random intercepts, Gaussian and t_3 data generating scenario. Average time to convergence (in seconds) for a chosen value of $G = 2$ with $r = 1$ and $G = 4$ with $r = 5$ and different quantiles.

Model	$q = 0.25$				$q = 0.50$			
	$r = 1$		$r = 5$		$r = 1$		$r = 5$	
	$n = 100$	$n = 250$	$n = 100$	$n = 250$	$n = 100$	$n = 250$	$n = 100$	$n = 250$
	Gaussian distribution							
Mixed	-	-	-	-	0.001	0.001	0.026	0.038
MixedNP	-	-	-	-	0.108	0.301	0.298	0.895
QRRE	0.042	0.262	0.146	0.443	0.020	0.092	0.050	0.164
FMMQ$_E$	1.203	2.708	1.513	4.709	1.577	2.784	1.286	3.200
FMMQ$_R$	0.525	1.024	0.148	0.276	0.814	1.369	1.073	1.243
FMQR	0.302	0.735	0.674	2.663	0.374	0.530	0.537	1.844
MQ	0.001	0.002	0.003	0.003	0.001	0.002	0.002	0.003
QR	0.001	0.002	0.002	0.003	0.001	0.002	0.002	0.003
	t_3 distribution							
Mixed	-	-	-	-	0.002	0.001	0.030	0.038
MixedNP	-	-	-	-	0.089	0.171	0.129	0.427
QRRE	0.061	0.296	0.146	0.252	0.022	0.100	0.135	0.374
FMMQ$_E$	0.938	1.789	0.758	0.775	0.961	1.180	0.526	1.256
FMMQ$_R$	0.547	1.332	0.828	1.034	0.788	1.296	0.367	0.758
FMQR	0.774	0.875	0.407	0.604	0.488	0.916	0.261	0.885
MQ	0.002	0.002	0.003	0.003	0.002	0.002	0.002	0.002
QR	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002

and their interaction as covariates. Figure 1 shows the normal probability plots for level 1 residuals and level 2 residuals (that is empirical Bayes estimates of individual-specific random effects): as it could be noticed, both indicate the presence of potentially influential observations in the data, with a number of large residuals. This finding is consistent with the Cook's distance plot for individual-specific random effects which is reported in Figure 2; the plot clearly suggests that some outliers are present in the analysed data. Therefore, using an M-quantile model with a bounded influence function seems quite a reasonable choice for these data.

We have analysed these data using the Finite Mixture of Quantile and M-quantile regression models at $q = 0.25, 0.5, 0.75$. The model specification includes fixed effects for the treatment group, time and the interaction between time and treatment group. The random part of the model includes a subject-specific intercept. As before, estimation for the M-quantile models is performed using the Huber influence function with $c = 1.345$ (FMMQ$_R$) and $c = 100$ (FMMQ$_E$).

Conditional on the value for q, the proposed model has been fitted for different values of the number of components in the finite mixture G; the log-likelihood value, the number of parameters, the penalised likelihood criteria and the time to convergence are reported in Table 8. The time to convergence of the EM algorithm, presented in Section 3, refers to an optimised R code on a laptop with a 2 GHz Intel Core i7 and 16 Gb RAM. The average time to convergence obviously increases with G, but it looks reasonable in all cases. AIC values lead to select a finite mixture model with $G \geqslant 3$ for all quantiles, but for FMQR and $q = 0.25$, where the best fitting model occurs for $G = 2$.
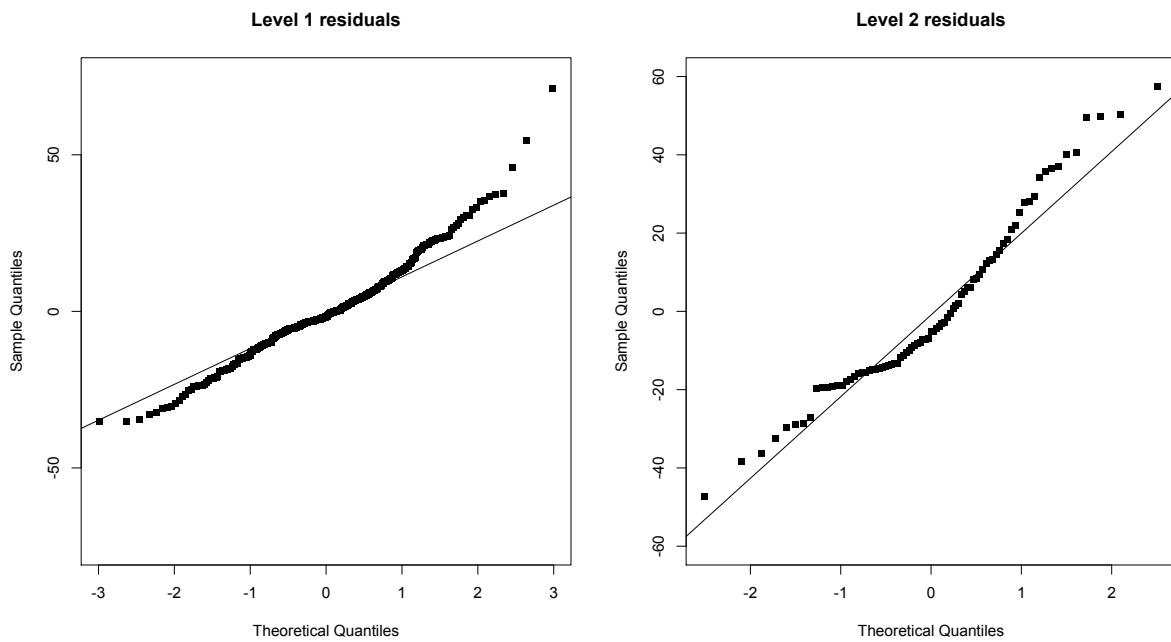
Figure 1: Pain Labora Data. Linear Mixed Effect Model. Normal probability plot for level 1 (left plot) and level 2 (random effects, right plot) residuals .
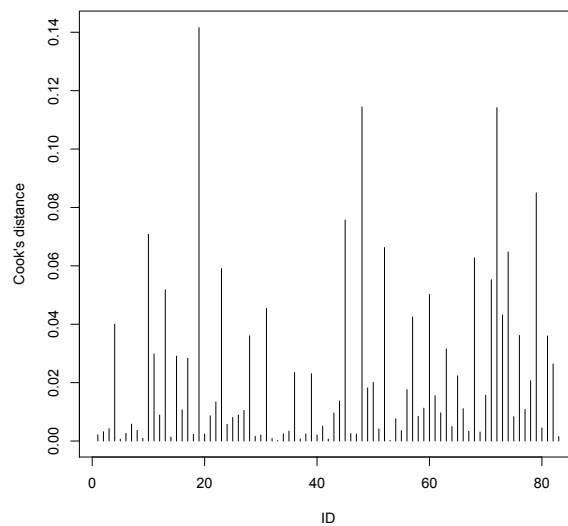

Figure 2: Model fit diagnostics for a Mixed fit to the Pain Labor data: unit level Cook's distances.

Table 9 shows the information on model fit for $FMMQ_E$, $FMMQ_R$, FMQR, with the number of components G chosen, for each model, according to the lowest value of AIC, and QRRE at different quantiles. At $q = 0.5$ the results for Mixed and MixedNP models have been also reported.

Table 10 displays the model parameter estimates and the corresponding standard errors. For $FMMQ_E$ and $FMMQ_R$, the estimates of the standard errors are obtained by the sandwich estimator and the bootstrap procedure using $B = 1000$ replications (results are in parenthesis); the bootstrap procedure has been used also for FMQR. As it can be noticed, the estimates for Treatment are not significant, but at $q = 0.25$ for QRRE and $FMMQ_E$, where the Treatment becomes slightly significant. The effect of the variable Time is significant: pain increases as time passes; also significant and worth of consideration is the interaction between Treatment and Time. We may guess that in the lower part of the distribution women have low pain, and hence can have only limited benefits from the treatment. On the other hand, higher quantiles are more influenced by the interaction between Treatment and Time, which is expected to decrease pain by around 10 mm. This effect is not evident for $FMMQ_E$ because it is affected by outliers. The effect of the interaction term is confirmed by looking at Figure 3 that shows a summary of the results for $FMMQ_R$ and FMQR; as expected, the placebo group experiences higher levels of pain than the treatment group. $FMMQ_R$ and FMQR help us to describe a meaningful picture of the distribution of labor pain while taking into account the longitudinal nature of the data. Overall, the treatment proves to be effective as women in the placebo group experience higher levels of pain than women in the treatment group; as time passes, the increase in the level of pain is higher for the placebo group than the treatment group. Examining Figure 3 closer, we notice that, for the treatment group at $q = 0.25$, the amount of pain experienced is almost constant. The same is true at the median, while at $q = 0.75$ the level of pain appears to be slightly increasing after two hours. This implies that the treatment is clearly effective for women that are more tolerant to pain, while its effect slightly decreases with time when women start to face some difficulties in staying tolerant to pain. By looking at Table 11, we may have some additional information on the estimated random intercept distribution; in all cases, the estimated distribution at $q = 0.25$ seems to be less diffuse than those estimated at $q = 0.50$ and $q = 0.75$, but for $FMMQ_R$, where no substantial changes can be observed. This may be due to higher-valued components being visible in the centre and the right tail of the distribution, but not in the left tail.

Table 8: Pain Labor Data. Log-likelihood, number of parameters, penalised likelihood criteria and time to convergence for different values of G and different quantiles.

| | FMMQ$_{\text{E}}$ | | | | | FMMQ$_{\text{R}}$ | | | | | FMQR | | | | |
| | G | | | | | G | | | | | G | | | | |
	2	3	4	5	6	2	3	4	5	6	2	3	4	5	6
$q = 0.25$															
log-lik	-1613.66	-1.583.73	-1578.04	-1583.02	-1577.746	-1608.10	-1565.18	-1584.26	-1565.96	-1568.09	-1583.34	-1583.43	-1583.34	-1583.34	-1583.34
# par.	7	9	11	13	15	7	9	11	13	15	7	9	11	13	15
BIC	3262.60	3214.49	3214.86	3236.583	3237.78	3251.48	3177.38	3227.31	3202.45	3218.47	3201.95	3213.71	3225.46	3237.22	3248.97
AIC	3239.33	3183.47	3176.08	3190.05	3183.49	3228.21	3146.36	3188.53	3155.92	3164.18	3178.68	3182.68	3186.68	3190.68	3194.68
Time	2.73s	3.72s	3.82s	3.92s	18.09s	1.59s	2.84s	2.35s	3.75s	3.98s	1.14s	0.32s	0.38s	0.45s	1.90s
$q = 0.50$															
log-lik	-1625.05	-1588.46	-1588.32	-1586.88	-1586.90	-1617.64	-1578.66	-1593.63	-1577.97	1577.70	-1602.26	-1595.19	-1595.19	-1555.76	-1555.49
# par.	7	9	11	13	15	7	9	11	13	15	7	9	11	13	15
BIC	3285.36	3223.95	3235.42	3244.31	3256.06	3270.56	3204.34	3246.04	3226.48	3237.69	3239.80	3237.41	3249.17	3182.05	3193.28
AIC	3262.09	3192.93	3196.64	3197.77	3201.77	3247.29	3173.32	3207.26	3179.95	3183.41	3216.53	3206.39	3210.39	3135.52	3138.99
Time	1.64s	2.08s	6.44s	10.22s	13.88s	1.21s	2.67s	2.84s	3.36s	4.58s	0.23s	0.42s	0.64s	0.96s	1.28s
$q = 0.75$															
log-lik	-1636.01	-1600.38	-1599.40	-1599.03	-1599.39	-1649.39	-1607.96	-1606.70	-1606.44	-1606.37	-1648.47	-1606.98	-1601.20	-1601.47	-1601.10
# par.	7	9	11	13	15	7	9	11	13	15	7	9	11	13	15
BIC	3307.29	3247.79	3257.57	3268.60	3281.08	3334.05	3262.94	3272.18	3283.41	3295.03	3332.21	3260.99	3261.31	3273.47	3284.49
AIC	3284.02	3216.77	3218.79	3222.06	3226.79	3310.79	3231.92	3233.40	3236.88	3240.74	3308.95	3229.96	3222.53	3226.94	3230.20
Time	1.48s	2.50s	12.16s	35.16s	8.60s	1.11s	2.94s	2.22s	8.01s	8.25s	0.36s	0.47s	0.90s	1.07s	0.84s

Table 9: Pain Labor Data. Log-likelihood, number of parameters, penalised likelihood criteria and time to convergence for chosen values of G and different quantiles.

Model	log-lik	# par.	BIC	AIC	Time
q=0.25					
QRRE	-1576.00	6	3187.29	3164.00	0.04s
FMMQ$_E$	-1578.04	11	3214.86	3176.08	3.82s
FMMQ$_R$	-1565.18	9	3177.38	3146.36	2.84s
FMQR	-1583.34	7	3201.95	3178.68	1.14s
q=0.50					
Mixed	-1595.56	6	3226.39	3203.12	0.04s
MixedNP	-1586.90	13	3191.77	3201.80	1.85s
QRRE	-1589.00	6	3213.64	3190.37	0.05s
FMMQ$_E$	-1588.46	9	3223.95	3192.93	2.08s
FMMQ$_R$	-1578.66	9	3204.34	3173.32	2.67s
FMQR	-1555.76	13	3182.05	3135.52	0.96s
q=0.75					
QRRE	-1589.00	6	3213.64	3190.37	0.04s
FMMQ$_E$	-1600.38	9	3247.79	3216.77	2.50s
FMMQ$_R$	-1607.96	9	3262.94	3231.92	2.94s
FMQR	-1601.20	11	3261.31	3222.53	0.90s

6.2 Case Study 2: Data on treatment of lead-exposed children

In this section, we apply the proposed modelling approach to a dataset from a placebo-controlled, double-blind, randomised trial to study the effect of succimer (a chelating agent) on children with blood lead levels (BLL) of 20–44 $\mu g/dL$. This trial study was conducted by the Treatment of Lead-Exposed Children (TLC) Trial Group (2000) in order to assess whether treating children with blood lead levels $< 45\mu g/dL$ is beneficial and can be done with acceptable safety. In fact, although lead encephalopathy had virtually disappeared from the United States, thousands of children still had sufficient lead exposure to produce cognitive impairment. Note that peak BLL occur at 20 to 30 months of age and are found to be inversely associated with cognitive test scores measured at ages 4 to 10 years. The trial was then designed to test the hypothesis that children with moderate BLL who were given succimer would have higher scores than children given placebo on a range of tests measuring cognition, behavior, and neuropsychological function three years after treatment.

Children in the trial were aged 12–33 months and they lived in deteriorating inner city housing. Placebo-treated children had a gradual decrease in blood lead level (see Fig. 1 in Treatment of Lead-Exposed Children (TLC) Trial Group, 2000). Succimer-treated children had an abrupt drop in blood lead level, much below that of the placebo group, followed by a rebound. Forty weeks after treatment, BLL were comparable in the two groups. Succimer was then found to lower blood lead level with few side effects. The dataset analysed here is available at `http://www.hsph.harvard.edu/fitzmaur/`

Table 10: Pain Labor Data. Parameter estimates and standard errors computed by sandwich estimator. Standard errors in parenthesis are obtained by using 1000 bootstrap replications.

Model	Variable	q=0.25		q=0.50		q=0.75	
		Estimate	Std. Error	Estimate	Std. Error	Estimate	Std. Error
Mixed	Treatement	-	-	-11.15	6.03	-	-
	Time	-	-	12.02	0.86	-	-
	Treatement:Time	-	-	-9.74	1.18	-	-
	σ_b	-	-	24.22	-	-	-
	scale	-	-	16.44	-	-	-
MixedNP	Treatement	-	-	7.64	1.28	-	-
	Time	-	-	11.22	0.35	-	-
	Treatement:Time	-	-	-9.05	0.47	-	-
	σ_b	-	-	25.85	-	-	-
	scale	-	-	16.72	-	-	-
QRRE	Treatement	-15.28	6.24	-7.63	6.47	-10.57	6.00
	Time	9.03	2.73	12.58	2.18	13.14	2.28
	Treatement:Time	-9.03	2.81	-12.10	2.44	-11.97	2.94
	σ_b	15.78		18.26		18.35	
	scale	18.23		17.16		20.27	
FMMQ$_E$	Treatement	-7.95	4.72 (4.32)	5.81	4.31 (6.43)	4.44	4.27 (8.10)
	Time	11.07	1.32 (1.79)	11.35	1.57 (1.53)	11.72	1.46 (1.52)
	Treatement:Time	-9.90	1.48 (1.95)	-9.23	1.88 (1.93)	-9.16	1.89 (2.20)
	σ_b	21.82		25.14		26.20	
	scale	14.69		16.86		15.56	
FMMQ$_R$	Treatement	2.09	3.09 (3.81)	2.48	3.74 (6.88)	-0.56	4.72 (9.75)
	Time	9.48	1.86 (1.52)	11.21	2.19 (2.58)	11.91	2.31(1.86)
	Treatement:Time	-7.92	1.84 (2.72)	-9.00	2.33 (2.38)	-9.47	2.40 (2.50)
	σ_b	23.34		24.22		25.32	
	scale	10.75		13.58		12.06	
FMQR	Treatement	2.00	4.48	0.49	2.24	-3.50	9.90
	Time	9.00	3.03	9.24	2.89	14.25	2.93
	Treatement:Time	-9.01	3.06	-8.74	2.93	-11.74	3.21
	σ_b	17.15		25.58		26.01	
	scale	21.65		16.18		21.34	

`ala/tlc.txt` and includes a subset of the original data for $n = 100$ children, 50 randomly assigned to treatment (succimer) and 50 to placebo group. For each child, BLL is available at baseline (or week 0), week 1, week 4 and week 6. This dataset has been reanalysed in Geraci and Bottai (2014).

Geraci and Bottai (2014) show a selection of individual BLL trajectories: for most children in the treatment group, a "U"-shaped curve describes the observed pattern, with a rebound between 1 to 4 weeks of treatment with succimer. On the other hand, a monotonic descending pattern can be mostly observed in the placebo group. The drop in BLL after one week of treatment is much larger in the treatment group than the slow decrease found in the placebo group. Given the observed "U"-shape, Geraci and Bottai (2014) decided to model the temporal trajectories using natural cubic splines with three degrees of freedom and knots placed at 0, 4 and 6 weeks. The B-spline basis has, therefore, two components: `T1` and `T2`, say. Here, we also consider a simpler quadratic model to handle the observed patterns: in the following we will denote the linear and the quadratic functions of time by `time` and `time2`.

Table 12 shows AIC values for a set of alternative models with several combinations of fixed and subject-specific random parameters. Binary variable `treat` takes value 1 for children treated with succimer and 0 for children treated with placebo. Results are displayed for models Mixed, QRRE, FMQR and FMMQ$_R$. The latter finite mixture models

Table 11: Pain Labor Data. Estimated random effect distribution at different quantiles.

	ζ_1	ζ_2	ζ_3	ζ_4
q=0.25				
FMMQ$_E$	-16.14	8.41	28.01	57.70
FMMQ$_R$	-0.02	31.05	65.24	
FMQR	-2.00	35.50		
q=0.50				
FMMQ$_E$	-4.21	26.31	67.63	
FMMQ$_R$	0.74	30.85	68.13	
FMQR	-0.49	8.50	35.50	72.25
q=0.75				
FMMQ$_E$	0.31	35.22	74.21	
FMMQ$_R$	3.75	34.63	68.98	
FMQR	5.00	38.50	71.50	
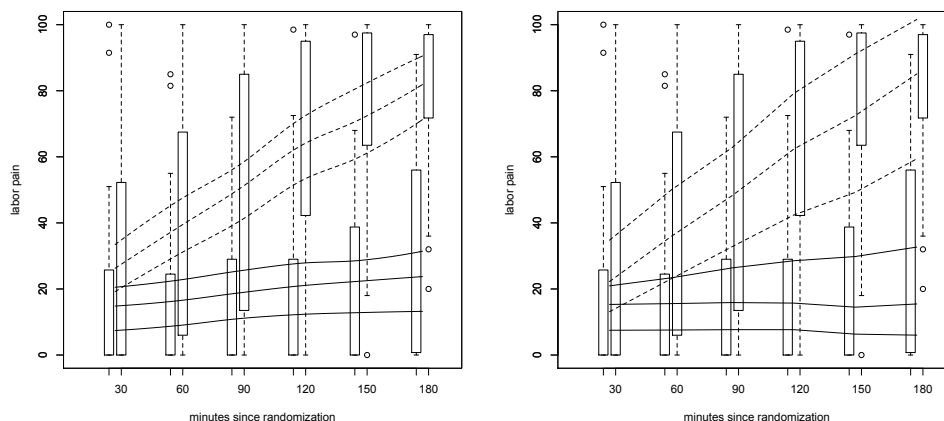
Figure 3: Boxplot of labor pain score. The lines represent the estimate of the quartile functions for the placebo group (solid) and the pain medication group (dashed) obtained by M-quantile model (left plot) and Quantile regression model (right plot).

have been fitted at $q = 0.50$ with a value of G ranging from two to ten; the resulting smallest value for AIC is reported, together with the chosen G. Model M3 is the one used in Geraci and Bottai (2014). The flexibility of spline-based (M1–M6) and of quadratic-in-time models (M7–M9) is required by the data and provides a better fit than a simple linear specification (M10–M11). Nonetheless, simpler quadratic models (M7–M9) achieve a better fit than spline based ones for most approaches. Among models M7–M9, it can be noted that the quadratic term is indeed needed only for treated units (M8–M9 vs. M7) and this is in line with the visual inspection of individual BLL trajectories. The best model in terms of AIC for Mixed, QRRE and FMQR is M9, and this is also a good compromise for FMMQ$_R$.

When M9 is fitted at $q = (0.25, 0.75)$, AIC selects 3 and 4 groups for FMQR, and 5 and 4 groups for FMMQ$_R$, respectively. Table 13 reports, for FMQR and FMMQ$_R$, the

Table 12: TLC data. AIC values for Mixed, QRRE, FMQR and FMMQ_R for alternative models with different combinations of fixed and subject-specific random variables. The intercept is always included in both the fixed and the random part of the model. FMQR and FMMQ_R have been fitted for $q = 0.50$ with a value for G ranging from 2 to 10; the resulting smallest value for AIC is reported with corresponding value of G.

Mod.	Fixed	Random	Mixed	QRRE	FMMQ_R	G	FMQR	G
M1	`treat * (T1 + T2)`	`treat * (T1 + T2)`	2586.3	2538.3	2512.1	4	2497.5	6
M2	`treat * (T1 + T2)`	`treat * T1 + T2`	2592.2	2526.7	2509.2	4	2517.9	4
M3	`treat * (T1 + T2)`	`treat * T1`	2584.2	2517.2	2517.5	6	2516.8	4
M4	`treat * (T1 + T2)`	`treat + T1 + T2`	2585.8	2522.6	2518.6	4	2522.5	4
M5	`treat * (T1 + T2)`	`treat * T2`	2581.2	2533.9	2509.9	4	2532.0	4
M6	`treat * (T1 + T2)`	`treat`	2582.0	2525.7	2542.2	4	2528.3	4
M7	`treat * (time + time2)`	`treat * time`	2576.2	2517.6	2541.0	3	2495.6	4
M8	`treat * time + treat:time2`	`treat * time`	2572.3	2515.3	2547.1	4	<u>2495.1</u>	4
M9	`treat * time + treat:time2`	`treat + time`	<u>2569.1</u>	<u>2507.7</u>	<u>2482.6</u>	4	2498.6	4
M10	`treat * time`	`treat * time`	err.	err.	2610.2	4	2635.1	3
M11	`treat * time`	`treat + time`	2687.0	2636.9	2633.6	3	2633.8	3

values of the entropy in posterior classification and the number of units classified in each group for each quantile. Classification is performed by using, for each unit, the maximum a posteriori probability, i.e. $\max_{k=1}^{G} \tau_{ik,q}$. For a given quantile, entropy is measured, within each individual, by taking the mean of the following quantity, $-\sum_{k=1}^{G} \tau_{ik,q} \log(\tau_{ik,q})$. Small values of such a measure denote groups with well clustered units, while relatively larger values denote groups with more heterogeneous units. By looking at the values in Table 13, at the estimated trajectories for treatment and control units within each group and at standard level estimates, we note that for $q = 0.25$, FMMQ_R could be better fitted using 4 groups and, for $q = 0.75$, both FMQR and FMMQ_R could be better fitted using 3 groups. We have thus decided to consider this more parsimonious models. Note that AIC values, not reported here, support this choice taking only slightly larger values under these more parsimonious settings.

Figure 4 shows for FMQR and FMMQ_R the estimated trajectories for $q = 0.25, 0.50, 0.75$ separately for treatment and control units. In the legend of each panel, the number of units in each group is also reported. Groups are colour coded and well separated. The classification looks similar in FMMQ_R and FMQR: a difference can be noted for $q = 0.25$ where FMMQ_R isolates a group of children with relatively large values of BLL.

The estimated regression quantiles and M-quantiles, together with their 95% confidence intervals using error bars are reported in Figure 5. Standard errors have been computed using 1,000 bootstrap replicates for FMMQ_R and FMQR, and using 100 bootstrap replicates for QRRE. Estimates and 95% confidence intervals from the Mixed fit have been reported for comparison as well. The intercepts can be interpreted as the quantiles of the conditional distribution of BLL in untreated children at week 0. Consistently with the findings of Geraci and Bottai (2014), the estimated treatment effect shows that the BLL distribution in the group of treated and untreated children differs only at lower quantiles. The negative coefficient for time provides evidence of a fall of BLL in the placebo group

Table 13: TLC data. Entropy and number of units for each group for FMQR and FMMQ$_R$ at $q = (0.25, 0.50, 0.75)$. For each model, the maximum Entropy value $-\log(G)$ – is also reported.

q	k	FMQR Entropy	n_k	k	FMMQ$_R$ Entropy	n_k
0.25	1	0.134	37	1	0.157	49
	2	0.392	37	2	0.779	25
	3	0.282	26	3	1.004	12
				4	0.866	8
				5	0.422	6
max(Entropy)	log(3)	1.099		log(5)	1.609	
0.50	1	0.366	38	1	0.264	36
	2	0.635	19	2	0.586	32
	3	0.453	25	3	0.267	27
	4	0.353	18	4	0.114	5
max(Entropy)	log(4)	1.386		log(4)	1.386	
0.75	1	0.686	28	1	0.836	30
	2	0.768	21	2	0.632	4
	3	0.377	31	3	0.440	47
	4	0.214	20	4	0.084	19
max(Entropy)	log(4)	1.386		log(4)	1.386	

over the time period at approximately uniform rate across the distribution (see also Fig. 4). In the treatment group, there is a fall of BLL levels at all quantiles but at faster rates than those in the placebo group (see the large negative estimates for `treatment:time` coefficients). The positive estimates for the `treatment:time2` coefficients reflect the U-shape that we observed in Fig. 4. The magnitude of this effect is weaker at higher levels of lead concentration.

7 Final remarks

The paper offers to the prospective data analyst some tools for modelling a general set of location parameters in the presence of data with clustered/hierarchical structure. In particular, we extend quantile and M-quantile regression to finite mixtures of quantile and M-quantile regression models, where components in the finite mixture represent clusters of individuals with homogeneous values of model parameters, conditional on a given quantile.

Simulation experiments show that the proposed models generally perform better than the competitors across the simulated scenarios, especially with respect to efficiency as the number of within-cluster measurements increases. The use of the methodology is facilitated by the availability of a computationally efficient algorithm developed in `R`. The models have been applied to reanalyse two well-known case studies: a longitudinal study on labour pain (Davis, 1991) and a placebo-controlled, double-blind, randomised
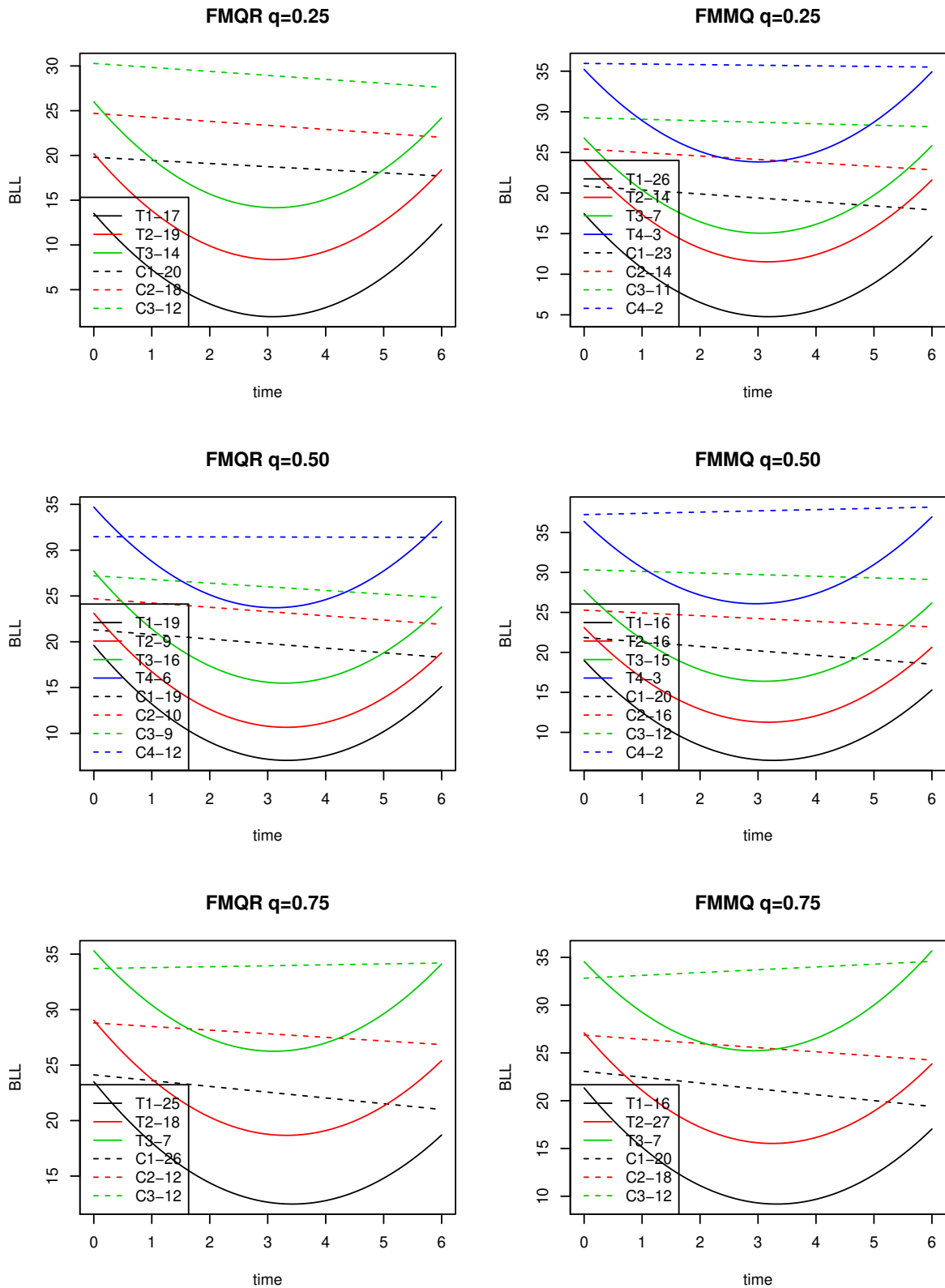
Figure 4: TLC data. Estimated trajectories for $\mathrm{FMMQ_R}$ and FMQR at $q = 0.25, 0.50, 0.75$ for each group identified with the mixtures and separately for the treatment (solid curves) and the control (dashed lines) group. Lines are colour coded according to the group and the number of units belonging to each category is reported in the legend.
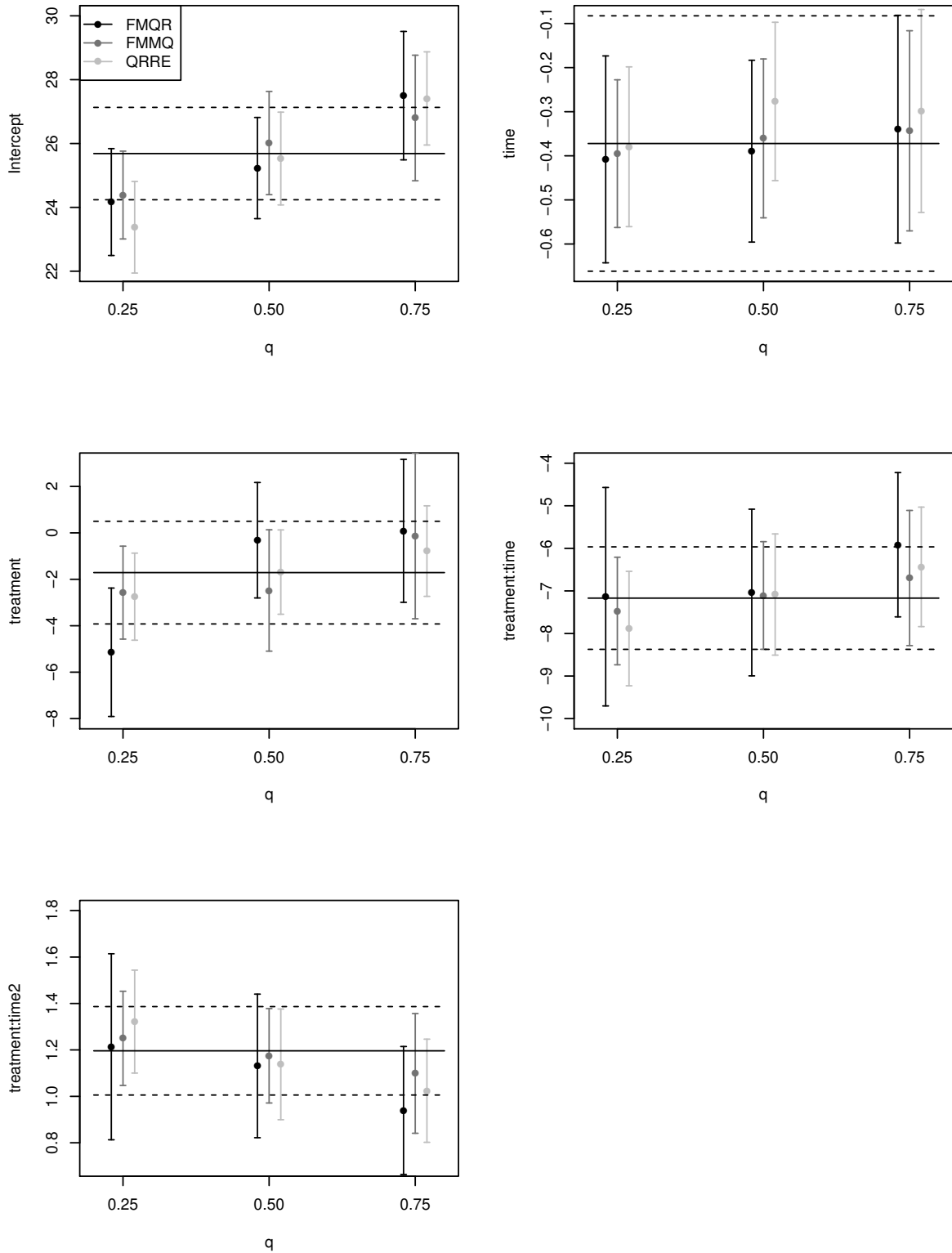
Figure 5: TLC data. Estimated coefficients for the fixed part of model M9 at quantiles $q = 0.25, 0.50, 0.75$ and 95% bootstrap-based confidence intervals (error bars) for FMMQ_R, FMQR and QRRE. Mixed estimates (black lines), using `lme` and 95% confidence intervals (dashed lines) are also reported for comparison.

35

trial entailing children with blood lead levels (BLL) of 20–44 $\mu g/dL$ (Treatment of Lead-Exposed Children (TLC) Trial Group, 2000). The results of the two studies show the capacity of the models to be adapted to different kinds of applications: random intercepts for the first case and general random coefficients in the second study.

The proposed techniques envision a whole modelling framework that extends in various directions existing statistical models. First of all, by suitably setting the tuning constant c to a relatively large value, it is possible to extend the classical finite mixture of linear regression models beyond the mean within the expectile framework. On the other hand, when $q = 0.5$, the choice of a relatively small value for the tuning constant c allows for a robust M-type extension of the classical finite mixture of linear regression models that can handle the presence of outlying observations. This robustification can also go beyond the mean and pertain the other expectiles of the conditional distribution of the response given the covariates. The important role of the tuning constant c also envisions new research towards the development of estimation strategies that allow for the joint estimation of its value and of model parameters.

The introduction of a nonparametric distribution for the random effects represents an important tool to further robustify M-quantile regression to model misspecification. Recently, Tzavidis et al. (2015) have proposed a new approach called M-quantile random effect regression, for modelling multilevel data. This methodology allows the estimation of a 2-level random intercept model. A line of research may be to compare the performance of this new method with a finite mixture of M-quantile models. **In addition, given the particularly valuable contribution of finite mixtures in generalised linear modelling, it will be interesting to extend them in an expectile and M-quantile fashion. Finally, extension to the multivariate case is also on the agenda, although the development could be hindered by the definition of expectiles and M-quantiles in more than one dimension.**

Acknowledgements

References

Abramowitz, M. and Stegun, I. (1964). *Handbook of Mathematical Functions.* National Bureau of standards, Washington DC.

Aitkin, M. (1996). A General Maximum Likelihood analysis of Overdispersion in Generalized Linear Models. *Statistics and Computing*, 6:127–130.

Aitkin, M. (1999). A General Maximum Likelihood analysis of Variance Components in Generalized Linear Models. *Biometrics*, 55:117–128.

Aitkin, M., Francis, B., and Hinde, J. (2005). *Statistical Modelling in GLIM*. Oxford University Press 2nd edition.

Alfó, M. and Trovato, G. (2004). Semiparametric mixture models for multivariate count data, with application. *Econometrics Journal*, 7:426–454.

Bianchi, A., Fabrizi, E., Salvati, N., and Tzavidis, N. (2015). M-quantile regression: diagnostics and parametric representation of the model. *Working paper*, available at `http://www.sp.unipg.it/surwey/dowload/publications/24-mq-diagn.html`.

Bianchi, A. and Salvati, N. (2015). Asymptotic properties and variance estimators of the M-quantile regression coefficients estimators. *Communications in Statistics - Theory and Methods*, 44:2416–2429.

Breckling, J. and Chambers, R. (1988). *M*-quantiles. *Biometrika*, 75:761–771.

Cantoni, E. and Ronchetti, E. (2001). Robust Inference for Generalized Linear Models. *Journal of the American Statistical Association*, 96:1022–1030.

Davis, C. (1991). Semi-parametric and non-parametric methods for the analysis of repeated measurements with applications to clinical trials. *Statistics in Medicine*, 10:1959–1980.

DeSarbo, W. and Cron, W. (1988). A maximum likelihood methodology for clusterwise regression. *Journal of Classification*, 5:249–282.

Farcomeni, A. (2012). Quantile regression for longitudinal data based on latent Markov subject-specific parameters. *Statistics and Computing*, 22:141–152.

Follmann, D. and Lambert, D. (1989). Generalizing logistic regression by nonparametric mixing. *Journal of the American Statistical Association*, 84:295–300.

Friedl, H. and Kauermann, G. (2000). Standard errors for EM estimates in generalized linear models with random effects. *Biometrics*, 56:761–767.

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

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

Geyer, C. and Thompson, E. (1992). Constrained Monte Carlo maximum likelihood for dependent data. *Journal of the Royal Statistical Society B*, 54:657–699.

Gueorguieva, R. (2001). A multivariate generalized linear mixed model for joint modelling of clustered outcomes in the exponential family. *Statistical Modelling*, 1:177–193.

Hennig, C. (2000). Identifiability of models for clusterwise linear regression. *Journal of Classification*, 17:273–296.

Huber, P. (1964). Robust estimation of a location parameter. *Ann. Math. Stat.*, 35:73–101.

Huber, P. (1973). Robust Regression: Asymptotics, Conjectures and Monte Carlo. *The Annals of Statistics*, 1:799–821.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pages 221–233. North Holland: Amsterdam.

Huber, P. J. (1981). *Robust Statistics*. John Wiley & Sons.

Jank, W. and Booth, J. (2003). Efficiency of Monte Carlo EM and Simulated Maximum Likelihood in Two-Stage Hierarchical Models. *Journal of Computational and Graphical Statistics*, 12:214–229.

Jones, M. C. (1994). Expectiles and m-quantiles are quantiles. *Statistics and Probability Letters*, 20:149–153.

Jung, S. (1996). Quasi-likelihood for median regression models. *Journal of the American Statistical Association*, 91:251–257.

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

Koenker, R. and Bassett, G. (1978). Regression quantiles. *Econometrica*, 46:33–50.

Koenker, R. and D'Orey, V. (1987). Computing regression quantiles. *Biometrika*, 93:255–268.

Kokic, P., Chambers, R., Breckling, J., and Beare, S. (1997). A measure of production performance. *Journal of Business and Economic Statistics*, 10:419–435.

Laird, N. M. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association*, 73:805–811.

Lipsitz, S., Fitzmaurice, G., Molenberghs, G., and Zhao, L. (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*, 46:463–476.

Liu, Q. and Pierce, D. (1994). A note on Gaussian-Hermite quadrature. *Biometrika*, 81:624–629.

Liu, Y. and Bottai, M. (2009). Mixed-effects models for conditional quantiles with longitudinal data. *Internation Journal of Biostatistics*, 5:1–22.

Louis, T. (1982). Finding the Observed information Matrix when Using the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 44:226–233.

McCulloch, C. (1994). Maximum likelihood estimation of variance components for binary data. *Journal of the American Statistical Association*, 89:330–335.

Munkin, M. K. and Trivedi, P. K. (1999). Simulated maximum likelihood estimation of multivariate mixed-Poisson regression models, with application. *The Econometrics Journal*, 2:29–48.

Newey, W. and Powell, J. (1987). Asymmetric least squares estimation and testing. *Econometrica*, 55:819–847.

Oakes, D. (1999). Direct Calculation of the Information Matrix via the EM Algorithm. *Journal of the Royal Statistical Society, series B*, 61:479–482.

Pinheiro, J. and Bates, D. (1995). Approximations to the Log-Likelihood Function in the Nonlinear Mixed-Effects Model. *Journal of Computational and Graphical Statistics*, 4:12–35.

Press, W., Teukolsky, S., Vetterling, W., and Flannery, B. (2007). *Numerical recipes: the art of scientific computing*. New York: Cambridge University Press.

Salvati, N., Tzavidis, N., Pratesi, M., and Chambers, R. (2012). Small area estimation via M-quantile geographically weighted regression. *TEST*, 21:1–28.

Street, J., Carroll, R., and Ruppert, D. (1988). A note on computing robust regression estimates via iteratively reweighed least squares. *The American Statistician*, 42:152–154.

Treatment of Lead-Exposed Children (TLC) Trial Group (2000). Safety and efficacy of succimer in toddlers with blood lead levels of 20-44 $\mu g/dl$. *Pediatr. Res.*, 48:593–599.

Tzavidis, N., Salvati, N., Schmid, T., Flouri, E., and Midouhas, E. (2015). Longitudinal analysis of the Strengths and Difficulties Questionnaire scores of the Millennium Cohort Study children in England using M-quantile random effects regression. *Journal of the Royal Statistical Society A*, in press.

Wang, P., Puterman, M., Cockburn, I., and Le, N. (1996). Mixed Poisson regression models with covariate dependent rates. *Biometrics*, 52:381–400.

Wang, Y., Lin, X., Zhu, M., and Bai, Z. (2007). Robust estimation using the huber funtion with a data-dependent tuning constant. *Journal of Computational and Graphical Statistics*, 16 (2):468–481.

Wedel, M. and DeSarbo, W. (1995). A mixture likelihood approach for generalized linear models. *Journal of Classification*, 12:21–55.

White, H. (1980). A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. *Econometrica*, 48:817–838.