Statistica Sinica **17**(2007), 929-943

MAXIMUM LIKELIHOOD INFERENCE IN ROBUST LINEAR MIXED-EFFECTS MODELS USING MULTIVARIATE t DISTRIBUTIONS

Peter X.-K. Song[1], Peng Zhang[2] and Annie Qu[3]

[1] *University of Waterloo,* [2] *University of Alberta and* [3] *Oregon State University*

Abstract: This paper focuses on the problem of maximum likelihood estimation in linear mixed-effects models where outliers or unduly large observations are present in clustered or longitudinal data. Multivariate t distributions are often imposed on either random effects and/or random errors to incorporate outliers. A powerful algorithm of maximum by parts (MBP) proposed by Song, Fan and Kalbfleisch (2005) is implemented to obtain maximum likelihood estimators when the likelihood is intractable. The computational efficiency of the MBP allows us to further apply a profile-likelihood technique for the estimation of the degrees of freedom in t-distributions. Comparison of the Akaike information criterion (AIC) among candidate models provides an objective criterion to determine whether outliers are influential on the quality of model fit. The proposed models and methods are illustrated through both simulation studies and data analysis examples, with comparison to the existing EM-algorithm.

Key words and phrases: AIC, EM-algorithm, Gauss-Newton algorithm, longitudinal data, MBP algorithm, outlier.

1. Introduction

In the context of maximum likelihood inference, a widely used robust approach to handling outliers or unduly large observations is to invoke heavy-tailed multivariate t distributions (Lange, Little and Taylor (1989)). In this paper we consider robust mixed modeling of clustered or longitudinal data using multivariate t distributions in the presence of influential outliers. A linear mixed-effects model (LMM) takes the form

$$y_i = X_i'\beta + Z_i'\alpha_i + \varepsilon_i,\ i = 1, \ldots, n, \tag{1.1}$$

where $y_i = (y_{i1}, \ldots, y_{im_i})'$ is the m_i-element response vector, $X_i = (x_{i1}, \ldots, x_{im_i})$ is a $p \times m_i$ matrix of covariates associated with the fixed effects, and $Z_i = (z_{i1}, \ldots, z_{im_i})$ is a $q \times m_i$ matrix of covariates associated with the random effects. Moreover, the random effects α_i are *i.i.d.* according to a q-dimensional density $p(\cdot|\eta)$ with parameter vector η, and ε_i are *i.i.d.* m_i-dimensional errors with density $p_i(\cdot|\sigma)$ whose mean is zero and variance-covariance matrix is $R(\sigma)$. The set

of parameters to be estimated is $\theta = (\beta, \eta, \sigma)$. Depending on where outliers may arise, they are called α-outliers if from the random effects, $p(\cdot|\eta)$, and ε-outliers if from the random errors, $p_i(\cdot|\sigma)$. In our model specification, either $p(\cdot|\eta)$, or $p_i(\cdot|\sigma)$, or both, may be assumed to be t-distributions.

Linear mixed models with non-normal random effects have drawn much attention recently in the literature. For example Pinheiro and Bates (2000) have provided an `nlme` library in R/Splus for fitting non-linear random effects models. The robustness of the LMM based on t-distributions has been discussed by Pinheiro, Liu and Wu (2001). Clearly, this class of models extends the popular normal random effects model because of the fact that a normal distribution is a special case of a t-distribution when the degrees of freedom is large. Since the t-distribution offers heavier tails than the normal distribution, the resulting model is often used to accommodate unduly large values that may arise either from sources of the random effects and/or the random errors.

Robust LMMs have also been studied in the literature by other researchers, such as Wakefield (1996). However, the utilization of such robust mixed-effects models is challenged by the difficulties of the maximum likelihood estimation (MLE) under the mixture of the normal and t distributions. The currently available solutions of obtaining the MLE in the literature include the EM algorithm and the Markov Chain Monte Carlo (MCMC) algorithm. Pinheiro et al. (2001) considered the t-t mixed-effects model in that they studied the EM-algorithm and other EM versions to accelerate the numeric convergence rate of MLE. Wakefield (1996) considered the t-normal mixed-effects model and established a Bayesian inference based on the MCMC algorithm. Both approaches are computationally intensive and require much analytical and numerical effort to implement. Moreover, the EM and the MCMC algorithm for the normal-t mixed-effects model have not been thoroughly studied, and are likely to be computationally intensive as well. Therefore, a unified algorithm that is simple and flexible enough to obtain the MLE under different combinations of t and normal distributions would be of great interest.

The purpose of this paper is to examine and utilize the new algorithm of maximization by parts (MBP) proposed by Song et al. (2005) for maximum likelihood inference in robust LMMs. The proposed procedure is simple, fast, and flexible for finding the MLE under various candidate LMMs for different distribution combinations. This fast numerical algorithm makes a joint comparison of many candidate models feasible. We provide a model selection procedure to assess candidate models by comparing the Akaike information (AIC) of each model. This procedure allows us to determine whether the outlying observations arise from the source of random effects and/or from the source of random errors, and is essential to detecting influential cases in robust data analysis. Such a thorough investigation has not yet been conducted in the literature.

This paper is organized as follows. Section 2 introduces the robust LMM and presents some basic discussion of maximum likelihood inference. Section 3 concerns the implementation of the BMP algorithm. Section 4 presents some numerical examples to illustrate the method, and Section 5 gives some concluding remarks. Some technical details are listed in the on-line supplementary document.

2. Preliminaries

2.1. Formulation

For ease of exposition, we take $m_i = m$ for $i = 1, \ldots, n$. Let $\theta = (\beta, \eta, \sigma)$ be the set of model parameters to be estimated.

We consider four possible LMMs: the normal-normal LMM in the absence of both α- and ε-outliers; the normal-t LMM in the presence of only ε-outliers; the t-normal LMM in the presence of only α-outliers; the t-t LMM in the presence of both α- and ε-outliers. To determine which LMM best fits the data, one can utilize a model selection procedure based on, for example, the Akaike information criterion (AIC). The AIC can be easily obtained if the maximum likelihood approach is applied.

For the classical normal-normal LMM, where both $p(\cdot|\eta)$ and $p(\cdot|\sigma)$ are normal in model (1.1), the likelihood function has a closed form and the related theory of the MLE has been studied extensively (e.g., McCulloch and Searle (2001)). The numerical implementation is available in many statistical software packages such as SAS PROC MIXED. Therefore, in the rest of this paper we focus on the problem of MLE in the other three types of LMMs.

According to Fang, Kotz and Ng (1990), the density of an r-dimensional t-distribution $Mt_r(d, D(\tau))$, with d degrees of freedom and a positive definite variance matrix $D(\tau)$, is

$$p(x|\tau) = \frac{\Gamma(\frac{d+r}{2})}{\{\pi(d-2)\}^{\frac{r}{2}}\Gamma(\frac{d}{2})}|D(\tau)|^{-\frac{1}{2}}\left\{1 + (d-2)^{-1}x^T D(\tau)^{-1}x\right\}^{-\frac{r+d}{2}}. \quad (2.1)$$

Note that the assumption of a finite variance (i.e., $d > 2$) is needed in order to ensure variance component parameters estimable in the setting of the mixed-effects models considered in this paper.

2.2. Likelihood functions

In model (1.1), suppose the random effects α_i follow a q-dimensional density $p(\alpha_i|\eta)$ and, conditionally on α_i the response vector y_i follows an m-dimensional density $p(y_i|\alpha_i, \theta)$. Then the likelihood function takes the form

$$L(\theta) = \prod_{i=1}^{n} p(y_i|\theta) = \prod_{i=1}^{n} \int_{\mathcal{R}^q} p(y_i|\alpha_i, \theta)p(\alpha_i|\eta)d\alpha_i. \quad (2.2)$$

The maximum likelihood estimation for θ can be obtained by maximizing the function $L(\theta)$ with respect to θ, which is done typically by solving the score equation $\dot{\ell}(\theta) = 0$, where $\dot{\ell}(\theta)$ is the first order derivative of the log-likelihood $\ln L(\theta)$. It is known that in general the related maximization procedure can be numerically difficult, because it requires the calculation of the q-dimensional integrals in (2.2) that have no closed form expression in the case of the multivariate t density $p(\cdot|\eta)$.

However, the t-t LMM is one exception, when the t distribution of the errors has the same degrees of freedom as that of the random effects. Following Pinheiro et al. (2001), given that $\alpha_i|\tau_i \sim N_q(0, D(\eta)/\tau_i)$, and that $\varepsilon_i|\alpha_i, \tau_i \sim N_m(0, \sigma I_m/\tau_i)$ and $\tau_i \sim \chi^2_d/d$, the resulting marginal density is m-dimensional t given by

$$p(y_i;\theta) = \frac{|V_i|^{-\frac{1}{2}}\Gamma(\frac{d+m}{2})}{[\Gamma(\frac{1}{2})]^m\Gamma(d/2)d^{m/2}}\left\{1 + \frac{(y_i - X_i\beta)'V_i^{-1}(y_i - X_i\beta)}{d}\right\}^{-\frac{d+m}{2}},$$

with $V_i = Z_iD(\eta)Z_i' + \sigma I_m$, where I_m is the m-dimensional identity matrix. It follows immediately that the log-likelihood of θ is given, subject to a constant, by

$$\ell(\theta) = -\frac{n}{2}\sum_{i=1}^{n}\ln|V_i| - \frac{d+m}{2}\sum_{i=1}^{n}\ln\left\{1 + \frac{(y_i - X_i\beta)'V_i^{-1}(y_i - X_i\beta)}{d}\right\}. \quad (2.3)$$

The likelihood function (2.2) of the normal-t LMM is the mixture of the m-variate normal distribution $N_m(X_i'\beta + Z_i'\alpha_i, R(\sigma))$ for the errors and the q-dimensional t distribution $Mt_q(d, D(\eta))$ for the random effects. The likelihood function (2.2) of the t-normal LMM can be similarly specified. In these two LMMs, numerical evaluation of the integrals is required in order to calculate their likelihood functions or the related derivatives. The MBP algorithm (Song et al. (2005)) has numerical efficiency and stability in solving the score equation, and only requires first order derivatives of the log-likelihood, which is very appealing. Therefore, it is of interest to examine the performance of the MBP algorithm that produces the MLE in all three types of LMMs.

3. Maximum Likelihood Estimation via MBP Algorithm

We choose the t-normal LMM to exemplify the implementation of the MBP algorithm for MLE. The MLE for the remaining two LMMs may be similarly carried out.

For the three models, we employ the method of profile likelihood to estimate the degrees of freedom d. That is, for each value of d in an interval $[3, B]$ with a suitably large constant B, $\ell(\theta|d)$ is maximized at the MLE $\widehat{\theta}(d)$. The estimator

of d is chosen such that the likelihood function is maximum on the interval $[0, B]$, namely

$$\widehat{d} = \arg \max_{d \in [3,B]} \ell(\widehat{\theta}(d)|d).$$

This profile likelihood approach is numerically feasible using a sequence of dense grid points on interval $[3, B]$, since the MBP algorithm converges to the MLE very quickly.

We now consider the MLE in the t-normal LMM. The MBP is by nature a fixed point algorithm, which requires one to specify a working model that resembles model (1.1) with t-distributed random effects. According to the information dominance principle (Song et al. (2005)), the closer the working model to the assumed model, the faster the MBP algorithm converges to the MLE. One obvious choice of the working model here is the normal-normal LMM, in which the distribution of the random effects is assumed to be normal $N_q(\mu, D)$. In particular, in order to achieve the desirable closeness, the variance matrix D is specified as being the same or close to that of the multivariate t. This matching of the first two moments ensures that the working model produces a consistent, although not fully efficient, estimator of θ. The utilization of the MBP algorithm for full likelihood inference in the t-normal LMM is established by the following derivation.

Let $\phi(\cdot)$ denote a normal density function. First, the likelihood function of the working normal-normal LMM is $L_w(\theta) = \prod_{i=1}^n \phi(y_i|\theta)$, where

$$\phi(y_i|\theta) = \frac{\phi(y_i, \alpha_i|\theta)}{\phi(\alpha_i|y_i, \theta)} = \frac{p(y_i|\alpha_i, \theta)\phi(\alpha_i|\theta)}{\phi(\alpha_i|y_i, \theta)}.$$

Since y_i is normally distributed conditional on the random effects α_i, it follows immediately that

$$p(y_i|\alpha_i, \theta) = \phi(y_i|\alpha_i, \theta) = \phi(y_i|\theta)\frac{\phi(\alpha_i|y_i, \theta)}{\phi(\alpha_i|\eta)}.$$

Entering this into (2.2), we can rewrite the likelihood function (2.2) as

$$L(\theta) = \prod_{i=1}^n \phi(y_i|\theta) \int \frac{p(\alpha_i|\eta)}{\phi(\alpha_i|\eta)}\phi(\alpha_i|y_i, \theta)d\alpha_i = L_w(\theta)L_e(\theta), \tag{3.1}$$

where $L_e(\theta) = \prod_{i=1}^n \int [p(\alpha_i|\eta)/\phi(\alpha_i|\eta)]\phi(\alpha_i|\mathbf{y}_i, \theta)d\alpha_i$. Thus, the log-likelihood is expressed as a sum of the working log-likelihood and the reminder log-likelihood $\ell(\theta) = \ell_w(\theta) + \ell_e(\theta)$, so the score equation is, $\dot{\ell}(\theta) = \dot{\ell}_w(\theta) + \dot{\ell}_e(\theta) = 0$. Given $\alpha_i \sim N_q(\mu, D(\eta))$, the working log-likelihood is given, subject to a constant, by

$$\ell_w(\theta) = -\frac{1}{2}\sum_{i=1}^n \log|\Sigma_i| - \frac{1}{2}\sum_{i=1}^n (y_i - X_i\beta - Z_i\mu)'\Sigma_i^{-1}(y_i - X_i\beta - Z_i\mu), \tag{3.2}$$

where $\Sigma_i = Z_i D(\eta) Z_i' + R(\sigma)$.

A merit of this likelihood partition is that the MBP will utilize the dominant piece ℓ_w to direct the search for the MLE, and the secondary piece ℓ_e only ensures the full efficiency of the estimator at convergence. Clearly, the likelihood ℓ_w of the working normal-normal LMM is analytically simple – we can easily obtain its second order derivatives. It is noted that in the remainder piece ℓ_e, the integral $\int [p(\alpha_i|\eta)/\phi(\alpha_i|\eta)]\phi(\alpha_i|\mathbf{y}_i, \theta) d\alpha_i$ can be viewed as essentially a weighted average discrepancy measure between the assumed and working distributions of the random effects under the working 'posterior' $\phi(\alpha_i|y_i, \theta)$ of the random effects. When the t-distribution has a large d, the ratio $p(\alpha_i|\eta)/\phi(\alpha_i|\eta)$ will be virtually 1, so that the ℓ_e becomes zero. Therefore, comparing the deviation of ℓ_e from zero for different candidate models will enable us to determine whether outliers arise from the random effects or from the random errors, and whether they are influential on the MLE, by assessing the difference in the estimation incurred from the removal of outliers. Furthermore, we can determine whether outliers have any impact on the quality of model fit via the AIC.

The integration evaluation in the ℓ_e is straightforward, because the $\phi(\alpha_i|y_i, \theta)$ is multivariate normal. When the dimension q is high, one may employ the Monte Carlo method; when q is low, one may instead apply the Gaussian-Hermite quadrature method (Liu and Pierce (1994)). In all our numerical examples, we adopt the quadrature method to evaluate related integrals. Note that the computational complexity of the two numerical integration methods is comparable.

3.2. The MBP algorithm

Suppose the degrees of freedom d is fixed in the t-normal LMM. To solve the score equation $\dot{\ell}(\theta) = 0$ without using the second order derivatives of the (complicated) log-likelihood ℓ, the MBP algorithm proceeds as follows:

Step 1: Acquire the consistent initial estimate $\theta^1 = (\beta^1, \eta^1, \sigma^1)$ by fitting the working normal-normal LMM.

Step k: Update to θ^k that solves $\dot{\ell}_w(\theta) = -\dot{\ell}_e(\theta^{k-1})$. Liao and Qaqish (2005) suggested a one-step Newton-Raphson update:

$$\theta^k = \theta^{k-1} - \{\ddot{\ell}_w(\theta^{k-1})\}^{-1} \dot{\ell}(\theta^{k-1}), \tag{3.3}$$

where $\ddot{\ell}_w(\theta^{k-1})$ is the Hessian matrix of the working model evaluated at the previous update θ^{k-1}. When this Hessian matrix is replaced by the corresponding minus Fisher Information of the working model, (3.3) becomes a one-step Fisher-scoring update. The Hessian matrix $\ddot{\ell}_w(\theta)$ of the working normal-normal LMM has been derived by many authors (for example, refer to Section 6.12 of McCulloch and Searle (2001, p.178)).

Iterate the above steps to convergence. According to Song et al. (2005), a simple way to verify the information dominance condition is to monitor the sequence of differences between two consecutive updates. If the differences diminish to zero over the iterations, this ensures the convergence of the MBP algorithm to the MLE.

The score vector $\dot{\ell}(\theta)$ is given as follows:

$$\dot{l}(\theta)=\sum_{i=1}^{n}\left\{\int p(y_i|\alpha_i)\phi(\alpha_i)d\alpha_i\right\}^{-1}\int\left\{\frac{\partial \ln p(y_i|\alpha_i)}{\partial\theta}+\frac{\partial \ln \phi(\alpha_i)}{\partial\theta}\right\}p(y_i|\alpha_i)\phi(\alpha_i)d\alpha_i,$$

where

$$\ln p(y_i|\alpha_i) = -2\ln\sigma - \frac{d+4}{2}\ln\left\{1+\frac{(y_i-X_i\beta-Z_i\alpha_i)'(y_i-X_i\beta-Z_i\alpha_i)}{(d-2)\sigma}\right\},$$
$$\ln \phi(\alpha_i) = -\frac{1}{2}\ln|D| - \frac{1}{2}\alpha_i' D^{-1}\alpha_i.$$

The first order derivatives of $\ln p(y_i|\alpha_i)$ and $\ln \phi(\alpha_i)$ with respect to the parameters are provided in Section 1.1 of the on-line supplementary document.

3.3. The MBP in other LMMs

The MBP algorithm in the normal-t LMM or the t-t LMM can be implemented similarly. Given the working normal-normal LMM, the true log-likelihood of the assumed model $\ell(\theta)$ is then decomposed in an additive form as $\ell_w(\theta)+\ell_e(\theta)$, where the reminder log-likelihood is given by the difference: $\ell_e(\theta) = \ell(\theta) - \ell_w(\theta)$. Again, the implementation of the MBP algorithm (3.3) in the normal-t LMM requires scores that are provided in Section 1.2 of the on-line supplementary document.

In the case of the normal-t LMM, since the distribution of the random effects, $p(\alpha_i|\eta)$, is normal, the full likelihood $L(\theta)$ in (2.2) and all the above scores can be directly evaluated by either the Gauss-Hermite quadrature method when dimension $q \leq 5$, or the Monte Carlo method when dimension $q > 5$.

In the case of the t-t LMM, even if the full likelihood function ℓ has a closed form expression (2.3), the MBP algorithm is still desirable because it avoids the derivation of second order derivatives, which may be tedious. The vector of scores $\dot{\ell}(\theta)$ with respect to the parameter $\theta = (\beta, \eta, \sigma)$ is given in Section 1.3 of the on-line supplementary document.

4. Simulation Experiment

This section presents a simulation study that aims to compare the MBP algorithm to the EM algorithm in the t-normal LMM. Pinheiro et al. (2001)

established the EM algorithm for the t-t LMM, which is different from the one we derive in this section. A brief description of the EM algorithm is provided below. The reason that we choose the t-normal mixture is that this model seems to be the most difficult one among the three models for both MBP and EM algorithms to handle. For simplicity, we only consider the case wherein the random intercepts are included.

4.1. EM algorithm

The EM algorithm for the t-normal LMM can be derived by expressing the model in a hierarchical form:

$$\begin{aligned} y_i|\alpha_i, b_i &\sim N_m(X_i\beta + \alpha_i \mathbf{1}_m, \sigma I_m),\ \alpha_i|b_i \sim N(0, \frac{\eta}{b_i}) \\ b_i &\sim \text{Gamma}(\frac{d}{2}, \frac{d}{2}), \end{aligned} \tag{4.1}$$

where $\mathbf{1}_m$ is the m-element vector of all elements being one, and $\text{Gamma}(\xi, \lambda)$ with the density function $p(b) = \lambda^{\xi} b^{\xi-1} \exp(-\lambda b)/\Gamma(\xi), b > 0, \xi > 0, \lambda > 0$. Integrating out the random effects α_i in (4.1) leads to the following simplified hierarchical representation:

$$y_i|b_i \sim N_m(X_i\beta, \sigma I_m + \frac{\eta}{b_i} J_m),\ b_i \sim \text{Gamma}(\frac{d}{2}, \frac{d}{2}), \tag{4.2}$$

where $J_m = \mathbf{1}_m \mathbf{1}_m^T$. See Section 2 of the on-line supplementary document for the proof of (4.2).

In the E-step, let $Q(\theta; \theta^k) = \text{E}\{\ell(\theta)|\theta^k, Y\}$, where θ^k denotes a known update value of the parameter from the previous iteration and Y denotes the entire set of observations. By simple algebra, we obtain

$$\begin{aligned} Q(\theta; \theta^k) \propto & -\frac{n(m-1)}{2} \ln \sigma \\ & -\frac{1}{2} \sum_{i=1}^{n} \text{E}\left\{ \ln\left(\sigma + \frac{m\eta}{b_i} \mid y_i, \theta^k \right) \right\} - \frac{1}{2\sigma} \sum_{i=1}^{n} \sum_{j=1}^{m} (y_{ij} - x_{ij}^T \beta)^2 \\ & + \frac{\eta}{2\sigma} \sum_{i=1}^{n} E\left(\frac{1}{\sigma b_i + m\eta} \mid y_i, \theta^k \right) \left\{ \sum_{j=1}^{m} (y_{ij} - x_{ij}^T \beta) \right\}^2, \end{aligned}$$

where the expectation is taken under the conditional distribution $p(b_i|y_i, \theta^k)$.

In the M-step, a new update value is obtained by maximizing the $Q(\theta, \theta^k)$, which may be performed by using the one-step Newton-Raphson algorithm (Rai and Matthews (1993)):

$$\theta^{k+1} = \theta^k - \left\{ \ddot{Q}(\theta; \theta^k) \right\}^{-1} \dot{Q}(\theta, \theta^k)_{|\theta=\theta^k}.$$

In this EM algorithm, the most time-consuming step is the evaluation of the conditional expectations involved in $\dot{Q}(\theta;\theta^k)$ and $\ddot{Q}(\theta;\theta^k)$. With no closed form expressions, they are evaluated by invoking the Monte Carlo method as suggested in Wei and Tanner (1990), which leads to the so-called Monte Carlo EM algorithm. Unfortunately, there are no simple formulas to generate a random number from the conditional distribution $p(b_i|y_i,\theta^k)$. Hence, we have to adopt the computationally intensive rejection-acceptance sampling algorithm (see details in Section 2 of the on-line supplementary document).

It is worth commenting that the approach of numerical evaluation for the expectations at the E-step is quite tedious. Difficulties arise from three aspects. (i) Because the conditional density $p(b_i|y_i,\theta)$ does not appear to have the form of $h(b)e^{-b^2}$, there does not exist a common set of quadrature points to carry out numerical integration collectively, as does the MBP algorithm. (ii) Each integral has its own integrand that requires a different choice of a finite interval. The different integrands will affect the decision on the number of knots, which divide the chosen finite interval into subintervals as required by, for example, Romberg's algorithm. (iii) The adaptive Gaussian quadrature method iteratively divides the chosen finite interval. In contrast, the Monte Carlo method generates one random sample useful to all integration evaluations, and it is also free of the finite interval constraint. All these differences lead to the popularity of the Monte Carlo EM algorithm.

4.2. Simulation results

We implement the MBP algorithm and a Gauss-Newton (GN) type algorithm (Ruppert (2005)) in this simulation study. Both algorithms require numerical evaluations of integrals. The Gauss-Hermite quadrature method is chosen for this task with 50 quadrature points at each evaluation. Table 1 lists the summary of the simulation results from both MBP and EM algorithms based on the t-normal LMM. The fixed effects were specified by one binary covariate (representing two types of treatments) with $n/2$ subjects being assigned to 1 and $n/2$ subjects assigned to 0. The true values of the parameters were $\beta_0 = 0.5$, $\beta_1 = 1.0$, $n = 100$, and $m = 5$. The random effects included only the random intercepts that were assumed to follow $t(3)$ with $\eta = 1$, and the random errors were assumed to be *i.i.d.* $N(0, 0.5^2)$ with $\sigma = 0.25$. Due to the extremely slow convergence of the EM algorithm, only 100 replicates were generated in this simulation. The naive estimate refers to the estimate obtained from the normal-normal LMM that has the same first two moments as those of the assumed t-normal LMM.

The simulation study provided the following result. (1) The MBP, GN and EM algorithms took almost the same number of iterations to convergence under

the convergence criterion: $\max_j |\theta_j^k - \theta_j^{k-1}| < 10^{-5}$. However, the MBP algorithm took only 23 seconds to complete 20 iterations, while it took 2,983 seconds (almost 1 hour) for the EM algorithm to complete 21 iterations. The contrast in computational efficiency between the two algorithms strongly favors the MBP algorithm. The MBP algorithm is also almost twice as fast as the GN algorithm. (2) The naive, MBP, GN and EM methods performed well in the estimation of of the fixed effects parameters and variance components. The results seem to be similar, and unbiased. (3) The asymptotic covariance matrix of the estimators was simply estimated by $(1/n)\sum \dot{\ell}_i(\hat{\theta})\dot{\ell}_i(\hat{\theta})'$. When n was small, this estimation did not seem to perform well, especially for the EM algorithm, based on the estimated standard errors. However, with only 100 replications this conclusion should be drawn with some reservation. The empirical standard errors of the estimates are provided in Table 1.

Table 1. Simulation comparison among the MBP, GN and EM algorithms based on the t-normal LMM over 100 replications. The $t(3)$ distribution is assumed for the random intercepts.

Parameter	Estimate	Iteration	Mean	Observed std. err.	CPU Time (in seconds)
β_0	Naive	0	0.4330	0.0865	-
	MBP	20	0.4782	0.1225	23.04
	EM	21	0.4988	0.1358	2983.5
	GN	15	0.4925	0.1162	51.01
β_1	Naive	0	1.0608	0.1342	-
	MBP	20	1.0098	0.1724	23.04
	EM	21	0.9988	0.1791	2983.5
	GN	15	0.9901	0.1688	51.01
η	Naive	0	1.0529	0.2573	-
	MBP	20	0.9511	0.2496	23.04
	EM	21	1.0338	0.2846	2983.5
	GN	15	1.0529	0.2632	51.01
σ	Naive	0	0.2475	0.0127	-
	MBP	20	0.2471	0.0096	23.04
	EM	21	0.2488	0.0125	2983.5
	GN	15	0.2627	0.0102	51.01

To examine the performance of the proposed four models in the presence of α-outliers, ε-outliers, and both α- and ε-outliers, we conducted further simulation studies that yielded similar findings to those given in Pinheiro et al. (2001), where only the comparison between the t-t LMM and normal-normal LMM was considered. A summary of the related results is presented in Section 3 of the on-line supplementary document.

5. Data Analysis

We re-analyze the orthodontic data analyzed previously by Pinheiro et al. (2001). These data were originally reported in an orthodontic study by Potthoff and Roy (1964). The measurements of the response variable included the distance from the pituitary gland to the pterygomaxillary fissure taken repeatedly at 8, 10, 12 and 14 years of age on a sample of 27 children, comprised of 16 boys and 11 girls. In the analysis by Pinheiro et al. (2001) two outliers were reported, namely boy M13 and boy M09 (respectively singled out in Figures 1 and 2). M13 was judged as an α-outlier arising from the random effects, suggested by the panels of individual intercepts and individual slopes in Figure 1. M09 was regarded as an ε-outlier arising from the random errors, indicated by the plot of standardized residuals obtained from individual regressions in Figure 2.

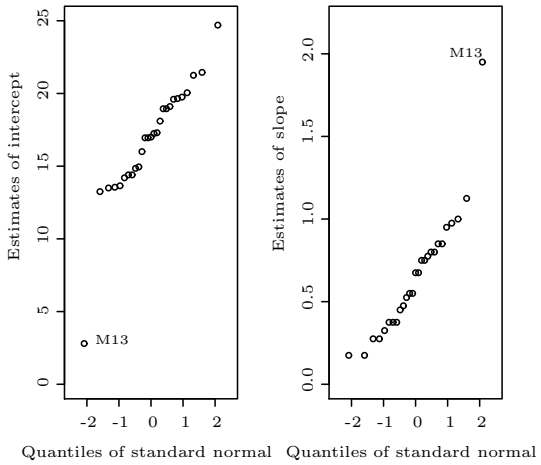

Figure 1. Diagnosis of outlier M13 through the plot of individual intercepts (left panel) and the plot of individual slopes (right panel) from a single subject regression analysis.

Now let y_{ij} bet the orthodontic distance for the ith subject at age t_j, and let $I_i(F)$ be an indicator variable for girls. Following Pinheiro et al. (2001), we consider the LMM of the form

$$\begin{aligned} y_{ij} = {} & \beta_0 + \beta_1 I_i(F) + \beta_2 t_j + \beta_3 I_i(F) \times t_j \\ & + \alpha_{0i} + \alpha_{1i} t_j + \varepsilon_{ij}, j = 1, \ldots, 4, i = 1, \ldots, 27, \end{aligned}$$

where $\alpha_i = (\alpha_{0i}, \alpha_{1i})$ denotes the vector of the random effects for the ith subject, and ε_{ij} are *i.i.d.* within-subject errors.

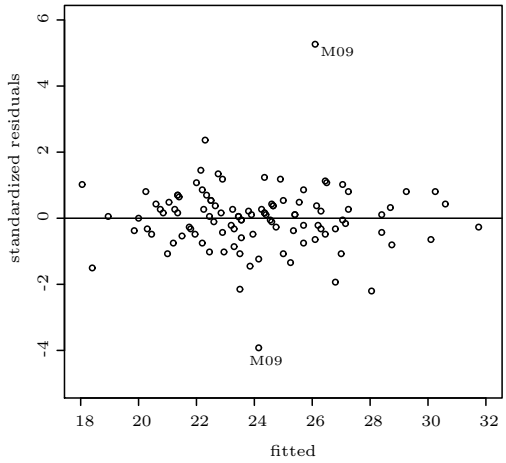

Figure 2. Diagnosis of outlier M09 through the plot of standardized residuals from a single subject regression analysis.

Compared to Pinheiro et al.'s analysis that assumed the t-t LMM, our analysis is more comprehensive, as we considered all four possible candidate LMMs with the mixtures of normal-normal, t-t, normal-t, and t-normal. The contrasts among the four models provide relevant evidence for us to determine outliers and their influence on the MLE using AIC. Note that only the t-t LMM was considered in Pinheiro et al.'s analysis, which could not assess whether certain outliers are influential because of the restriction of having equal degrees of freedom for both t distributions. Obviously, the estimated degrees of freedom will be small even if there is only one source of influential outliers.

Table 2 summarizes the estimates and standard errors of the fixed effects and the variance components under the four candidate LMMs via the MBP algorithm. The degrees of freedom, d, of the t distribution was estimated by the method of profile likelihood discussed in Section 3.

Since the difference of the AIC between the normal-normal and t-normal LMMs is marginal, the latter model accommodating the α-outlier M13 did not seem to improve the quality of fit much, although the actual estimates are slightly different. However, the difference of the AIC between the t-t and the normal-t is much smaller than that between the t-normal (or the normal-normal) and t-t (or normal-t). This indicates that accommodating the ε-outlier M09 appeared to have gained substantial improvement in the quality of fit. Relatively speaking, outlier M09 seems to be more troublesome or more influential than outlier M13 in the model selection. Based on the AIC, the normal-t LMM would provide

the best fit among the four candidate models. A formal test for the normal-normal versus normal-t is not trivial. Here we are testing $H_0 : d^{-1} = 0$ (infinite degrees of freedom) against $H_1 : d^{-1} > 0$. The null pertains to the parameter on the boundary of the parameter space. However, we can apply Self and Liang's (1987) mixture of chi-squares test. In this case, the observed likelihood ratio test statistic was found to be 17.605, which is greater than the critical value of 2.71 obtained from the 50:50 mixture of χ^2_0 (a point mass at 0) and χ^2_1 distributions at the significance level of 0.05. This suggests that the parameter d (degrees of freedom) in the t-distribution is finite.

Table 2. The results of orthodontic data analysis under four different linear mixed models.

	Model							
	N-N		t-N		t-t		N-t	
Parameter	Estimate	Std err	Estimate	Std err	Estimate	Std err	Estimate	Std err
β_0	16.3406	1.0186	16.2543	1.3276	16.9468	0.9051	17.1653	0.8560
β_1	1.0321	1.5958	1.3231	2.1085	0.6620	1.2688	0.3829	1.2064
β_2	0.7844	0.0860	0.7517	0.0456	0.7156	0.0762	0.7108	0.0719
β_3	−0.3048	0.1348	−0.2795	0.0934	−0.2567	0.1080	−0.2471	0.1003
σ	1.7162	0.2942	1.7151	0.0773	0.8880	0.2243	1.7301	0.7704
η_1	4.5569	0.9623	5.4781	0.6190	3.2755	2.9897	2.7273	1.2422
η_2	−0.1983	0.0957	0.0308	0.0318	−0.1336	0.2395	−0.0403	0.1382
η_3	0.0238	0.0185	0.0013	0.0008	0.0197	0.0219	0.0121	0.0146
d (DF)	–	–	4.2	–	5.0	–	3.8	–
Iteration	0	–	9	–	12	–	14	–
AIC	448.5816	–	446.4510	–	432.4600	–	428.2418	–

To further illustrate the robustness of the estimates in the ultimately prevailing normal-t LMM versus those in the naively selected normal-normal model, we fit both models again to the orthodontic data with the influential ε-outlier M09 removed. It was found that the estimated degrees of freedom increased by two units to 5.8 from 3.8. To compare the results obtained under a different data setting, we followed the DFBETA approach to influence analysis in classical regression analysis theory (Myers (1990)). That is, for each parameter we calculated the ratio of relative change,

$$RC(\theta_j) = \frac{|\theta^{\text{with}}_{j,N-N} - \theta^{\text{without}}_{j,N-N}|}{s.e.(\theta^{\text{without}}_{j,N-N})} \Big/ \frac{|\theta^{\text{with}}_{j,N-t} - \theta^{\text{without}}_{j,N-t}|}{s.e.(\theta^{\text{without}}_{j,N-t})},$$

between the two cases of with and without outlier M09 and between the normal-t and the normal-normal LMMs. A value of RC greater than 1 indicates that the

normal-normal LMM has a lower level of robustness than the the normal-t LMM. The results are listed in Table 3.

Table 3. The ratio of relative change of the maximum likelihood estimates for each parameter between the inclusion and exclusion of outlier M09. The comparison is made between the normal-normal and normal-t LMMs.

	Parameter							
θ_j	β_0	β_1	β_2	β_3	σ	η_1	η_2	η_3
RC(θ_j)	4.30	3.17	7.18	1.87	2.87	30.30	28.30	12.82

The estimates obtained by the normal-t LMM are evidently much more robust to the inclusion or exclusion of outlier M09 than those given by the normal-normal LMM. This demonstrates the importance of accommodating influential outliers via multivariate t-distributions in the analysis of longitudinal data by means of mixed-effects models.

6. Concluding Remarks

One issue that challenges the MBP algorithm is the large dimension of random effects q, say $q > 5$, when the numerical evaluation of integration becomes difficult. In practice, linear mixed-effects models with more than five random effects are not very common, although they are not impossible. When the case of high-dimensional random effects appears, one may invoke the t-t LMM that has a closed form expression of the likelihood, regardless of the dimension of random effects q. Again the MBP algorithm is recommended. However, for the t-N LMM and N-t LMM the MBP needs to invoke the Monte Carlo method to evaluate related integrals. More exploration is required to give satisfactory answers and useful practical guidelines with this regard. This will be reported in our future research.

The application of the MBP algorithm to other settings remains to be explored. In particular, it is unknown whether this algorithm may be helpful to resolve some difficulties of statistical inference in models for survival or time series data.

Acknowledgement

The authors are very grateful to an associate editor and the anonymous referee for constructive comments and valuable suggestions that led to an improvement for the paper. The first, second, and third author's research was supported by the Natural Sciences and Engineering Research Council of Canada, the start-up grant of University of Alberta, and the US National Science Foundation, respectively.

References

Fang, K.-T., Kotz, S. and Ng, K. W. (1990). *Symmetric Multivariate and Related Distributions.* Chapman and Hall, London.

Lange, K. L., Little, R. J. A. and Taylor, J. M. G. (1989). Robust statistical modeling using t-distribution. *J. Amer. Statist. Assoc.* **84**, 881-896.

Liao, J. G. and Qaqish, B. F. (2005). Discussion of maximization by parts in likelihood inference by Song, Fan, and Kalbfleisch. *J. Amer. Statis. Assoc.* **100**, 1160-1161.

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

Myers, R. (1990). *Classical and Modern Regression with Applications.* 2nd edition. Duxbury Press, Belmont.

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

Pinheiro, J. C., Liu, C. and Wu, Y. N. (2001). Efficient algorithms for robust estimation in linear mixed-effects models using the multivariate t distribution. *J. Comp. Graph. Statist.* **10**, 249-276.

McCulloch, C. E. and Searle, S. R. (2001). *Generalized, Linear, and Mixed Models.* Wiley, New York.

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

Rai, S. N. and Matthews, D. E. (1993). Improving the EM algorithm. *Biometrika* **49**, 587-591.

Ruppert, D. (2005). Discussion of maximization by parts in likelihood inference by Song, Fan, and Kalbfleisch. *J. Amer. Statis. Assoc.* **100**, 1161-1163.

Self, S. G. and Liang, K.-Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *J. Amer. Statist. Assoc.* **82**, 605-610.

Song, P. X.-K., Fan, Y. and Kalbfleisch, J. D. (2005). Maximization by parts in likelihood inference (with discussion). *J. Am. Statist. Assoc.* **100**, 1145-1167.

Wakefield, J. C. (1996). The Bayesian approach to population pharmacokinetic models. *J. Amer. Statist. Assoc.* **91**, 62-75.

Wei, G. C. G. and Tanner, M. A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithm. *J. Amer. Statist. Assoc.* **85**, 699-704.

Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1.

E-mail: song@uwaterloo.ca

Department of Mathematical and Statistical Sciences, University of Alberta, Edmonton, Alberta, Canada T6G 2G1.

E-mail: pengz@ualberta.ca

Department of Statistics, Oregon State University, Corvallis, Oregon 97331-4606, U.S.A.

E-mail: qu@science.oregonstate.edu

(Received March 2006; accepted September 2006)