Journal of Data Science **3**(2005), 415-438

Skew-normal Linear Mixed Models

R. B. Arellano-Valle[1], H. Bolfarine[2] and V. H. Lachos[2]
[1]*Pontificia Universidad Católica de Chile and* [2]*Universidade de São Paulo*

Abstract: Normality (symmetric) of the random effects and the within-subject errors is a routine assumptions for the linear mixed model, but it may be unrealistic, obscuring important features of among- and within-subjects variation. We relax this assumption by considering that the random effects and model errors follow a skew-normal distributions, which includes normality as a special case and provides flexibility in capturing a broad range of non-normal behavior. The marginal distribution for the observed quantity is derived which is expressed in closed form, so inference may be carried out using existing statistical software and standard optimization techniques. We also implement an EM type algorithm which seem to provide some advantages over a direct maximization of the likelihood. Results of simulation studies and applications to real data sets are reported.

Key words: EM algorithm, marginal likelihood, mixed effects model, skewness.

1. Introduction

The linear mixed model (LMM) have become a most commonly used for analyzing continuous repeated measures data from a sample of individuals in agricultural, environmental, biomedical, economical, and social science applications. Let $\mathbf{Y}_j$ be a $(n_j \times 1)$ vector of observed continuous responses for sample unit j, $j = 1, \ldots, m$. We assume that $\mathbf{Y}_j$ follows the general LMM:

$$\mathbf{Y}_j = \mathbf{X}_j\boldsymbol{\beta} + \mathbf{Z}_j\mathbf{b}_j + \boldsymbol{\epsilon}_j, \quad j = 1, \ldots, m\,, \tag{1.1}$$

where $\mathbf{X}_j$ of dimension $(n_j \times p)$ is the design matrix corresponding to the fixed effects, $\boldsymbol{\beta}$ of dimension $(p \times 1)$ is a vector of population-averaged regression coefficients called fixed effects, $\mathbf{Z}_j$ of dimension $(n_j \times q)$ is the design matrix corresponding to the $(q \times 1)$ random effects vector $\mathbf{b}_j$, and $\boldsymbol{\epsilon}_j$ of dimension $(n_j \times 1)$ is the vector of random errors. A standard but possibly restrictive assumption is that the random effects $\mathbf{b}_j$ and the residual components $\boldsymbol{\epsilon}_j$ are independent with

$$\mathbf{b}_j \overset{\text{iid}}{\sim} N_q(\mathbf{0}, \mathbf{D}), \quad \boldsymbol{\epsilon}_j \overset{\text{ind}}{\sim} N_{n_j}(\mathbf{0}, \boldsymbol{\psi}_j), \tag{1.2}$$

where $\mathbf{D} = \mathbf{D}(\boldsymbol{\alpha})$ and $\boldsymbol{\psi}_j = \boldsymbol{\psi}_j(\boldsymbol{\gamma})$, $j = 1, \ldots, m$, are dispersion matrices, usually associated with the variability among- and within-individuals, which are depending on unknown and reduced parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$, respectively. Frequently, estimation methods for the parameters in model (1.1)-(1.2) are maximum likelihood and restricted maximum likelihood (Lindstrom and Bates, 1988). Confidence intervals and hypothesis testing for the parameters are generally based on asymptotic results and softwares such as *SAS proc mixed* (Littell, Milliken, Stroup and Wolfinger, 1996) or *S-plus lme* (Pinheiro and Bates, 2000) incorporate procedures for analyzing LMM under this assumptions.

Though model (1.1)-(1.2) offers great flexibility for modelling these effects, it suffers from the same lack of robustness against departures from distributional assumptions as other statistical models based on the Gaussian distribution and may be too restrictive to provide an accurate representation of the structure that is present in repeated measures and clustered data. From a practical point of view, the most commonly adopted approach to achieve multivariate normality involves variables transformation. Although such methods may give reasonable empirical results, it should be avoided if a more suitable theoretical model can be found (Azzalini and Capitanio, 1999). Thus, considerable interest has focused on relaxing the normality assumption and jointly estimating the random effects and model parameters. In this context, recent proposals have been made based in replacing the assumption of normality by a weaker assumption that only the random effects have a "smooth" density that may be skewed. For example, Zhang and Davidian (2001) use the semiparametric (SP) density representation proposed by Gallant and Nychka (1987) to characterize the random effect density; an appealing feature of this approach is that the likelihood for all model parameters may be expressed in a closed form. Alternatively, Verbeke and Lesaffre (1996) adopt a mixture of normals representation and carry out inference via an EM algorithm (see Verbeke and Molemberghs, 2000, chap. 12). Magder and Zeger (1996) use a form of non-parametric maximum likelihood based normal densities and uses somewhat *ad hoc* fitting and assessment of the fit. Tao, Palta, Yandell, and Newton (1999) estimate the density of a scalar random effect via a predictive recursive algorithm. Recently, Sahu et al. (2003) proposed and alternative model suitable for Bayesian implementation, which seems not to be adequate for maximum likelihood implementation. We propose an alternative method that is particularly attractive for linear mixed models by assuming that both the random effect and the model errors follow a skew-normal distribution. Our approach may offer advantages of more efficient estimators and algorithms (for special cases) and also of practical interpretation for model parameters. It also has the advantage of providing readily available information matrices.

The plan of the paper is as follows. In Section 2, for the sake of completeness,

we consider a multivariate extension of the univariate skew-normal distribution proposed by Azzalini (1985). Properties like moments and stochastic representation of this multivariate distribution are also discussed. In Section 3 the skew-normal linear mixed model (SNLMM, hereafter) is defined extending the usual normal mixed model. The marginal density of $\mathbf{Y}_j$ is obtained analytically by integrating out the random effects $\mathbf{b}_j$, $j = 1, \ldots, m$, leading to the observed (marginal) likelihood function that can be maximized directly by using existing statistical softwares such as Ox, R or Matlab. We point out that the analytical expression for the likelihood function provided in this paper is not available elsewhere in the literature. Section 4 presents an EM type algorithm which presents advantages over the direct maximization approach, specially in terms of robustness with respect to starting values and a closed form to estimating $\boldsymbol{\beta}$ in the iterative process. Section 5 reports results of a simulation study and Section 6 reports applications to a real data set indicating the usefulness of the approach.

2. A Skew-normal Distribution

In this section we introduce the multivariate skew-normal distribution that will be used in defining the SNLMM considered in the following section. We start by giving an important notation that will be used through the whole paper and presenting a review of the bibliography in univariate skew-normal models.

Let $\phi_n(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\Phi_n(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ be the probability density function (pdf) and the cumulative distribution function (cdf), respectively, of the $N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution evaluated at $\mathbf{x}$. When $\boldsymbol{\mu} = \mathbf{0}$ and $\boldsymbol{\Sigma} = \mathbf{I}_n$ (the $n \times n$ identity matrix), we denote these functions as $\phi_n(\mathbf{x})$ and $\Phi_n(\mathbf{x})$.

As considered in Azzalini (1985), a random variable Y follows a univariate skew-normal distribution with location parameter μ, scale parameter σ^2 and skewness parameter λ if the pdf of Y is given by

$$f_Y(y) = 2\phi_1\left(y|\mu, \sigma^2\right) \Phi_1\left(\lambda \frac{y-\mu}{\sigma}\right). \tag{2.1}$$

Note that if $\lambda = 0$ then the density of Y in (2.1) reduces to the density of the normal distribution. We use the notation $Y \sim SN_1(\mu, \sigma^2, \lambda)$ to denote this distribution, which will be reduced to $Y \sim SN_1(\lambda)$ when $\mu = 0$ and $\sigma^2 = 1$. Properties of this distribution can be found in Azzalini (1985) and Henze (1986).

Studies on multivariate skew-normal distributions are considered in Azzalini and Dalla Valle (1996), Azzalini and Capitanio (1999), Branco and Dey (2001), Sahu et al. (2003), among others. Arellano-Valle, del Pino and San Martin (2002) show that many of the properties of the multivariate skew-normal distribution hold for a general class of skewed distributions obtained from a symmetric class, defined in terms of independence conditions on signs and absolute values and give

general formula to obtain skewed pdf's. From these results, Arellano-Valle and Genton (2005) introduce the class of fundamental skewed distributions, giving an unified approach to obtain multivariate skew distributions starting from symmetric ones. In this work, we consider a special case of the fundamental skew-normal distribution, proposed by Arellano-Valle and Genton (2005) (see also Azzalini and Dalla-Valle, 1996 and Azzalini and Capitanio, 1999). The definition is given in the following.

Definition 1: An n-dimensional random vector $\mathbf{Y}$ follows a skew-normal distribution with location vector $\boldsymbol{\mu} \in \mathbb{R}^n$, dispersion matrix $\boldsymbol{\Sigma}$ (a $n \times n$ positive definite matrix) and skewness vector $\boldsymbol{\lambda} \in \mathbb{R}^n$, if its pdf is given by

$$f_{\mathbf{Y}}(\mathbf{y}) = 2\phi_n(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma})\Phi_1(\boldsymbol{\lambda}^T\boldsymbol{\Sigma}^{-1/2}(\mathbf{y} - \boldsymbol{\mu})), \quad \mathbf{y} \in \mathbb{R}^n. \tag{2.2}$$

We denote this by $\mathbf{Y} \sim SN_n(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda})$ and by $\mathbf{Y} \sim SN_n(\boldsymbol{\lambda})$ when $\boldsymbol{\mu} = \mathbf{0}$ and $\boldsymbol{\Sigma} = \mathbf{I}_n$, the n-dimensional identity matrix.

Remark 1: Since the condition that $\Phi_1(-w) = 1 - \Phi_1(w)$ for all $w \in \mathbb{R}$ is sufficient to guarantee that (2.2) is a pdf, we can then use different reparameterizations to represent the asymmetric parameter $\boldsymbol{\lambda}$, as for example:

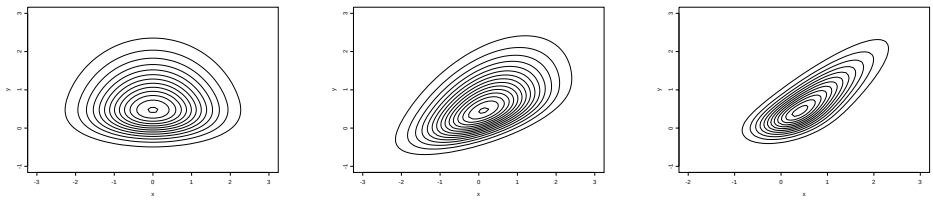
$$\boldsymbol{\lambda} = \frac{\boldsymbol{\Delta}^{-1/2}\boldsymbol{\delta}}{\sqrt{1 - \boldsymbol{\delta}^T\boldsymbol{\Delta}^{-1}\boldsymbol{\delta}}}, \tag{2.3}$$

for some $\boldsymbol{\delta} \in \mathbb{R}^n$ and positive definite $n \times n$ matrix $\boldsymbol{\Delta}$ such that $\boldsymbol{\delta}^T\boldsymbol{\Delta}^{-1}\boldsymbol{\delta} < 1$. Two special cases can be considered; $\boldsymbol{\Delta} = \boldsymbol{\Sigma}$, which is just the reparameterization used by Azzalini and Dalla-Valle (1996), and $\boldsymbol{\Delta} = \mathbf{I}_n$, which is used in Arellano-Valle and Genton (2005). In a more general way, we can replace in (2.2) the asymmetric part (or skewing function; see Genton and Loperfido, 2001) $\Phi_1(\cdot)$ by an arbitrary function $Q(\cdot)$ on $[0, 1]$, which depends on $\mathbf{y}$ trough an even real function (or antisymmetric function; see Arellano-Valle and del Pino, 2003), say $w(\mathbf{y})$, and is such that $Q(w(-\mathbf{y})) = Q(-w(\mathbf{y})) = 1 - Q(w(\mathbf{y}))$. Thus, the skew-normal distribution in (2.2) can be extended by considering

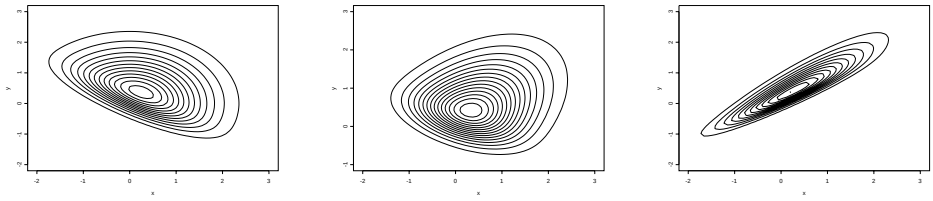
$$f_{\mathbf{Y}}(\mathbf{y}) = 2\phi_n(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma})Q(w(\mathbf{y})), \quad \mathbf{y} \in \mathbb{R}^n. \tag{2.4}$$

Many properties of the above skew-normal distribution may be derived from the results developed by Arellano-Valle and Genton (2005) (see also Arellano-Valle et al., 2002 and Arellano-Valle and del Pino, 2003). From there it follows, for example, the stochastic representation given next for an standardized skew-normal random vector.

(a) For $\boldsymbol{\lambda} = (0, 3)^T$ and $\rho = 0$, 0.5, 0.9, respectively.



(b) For $\boldsymbol{\lambda} = (2, 3)^T$ and $\rho = 0$, 0.5, 0.9, respectively.



(c) For $\boldsymbol{\lambda} = (-2, 2)^T$ and $\rho = 0$, 0.5, 0.9, respectively.

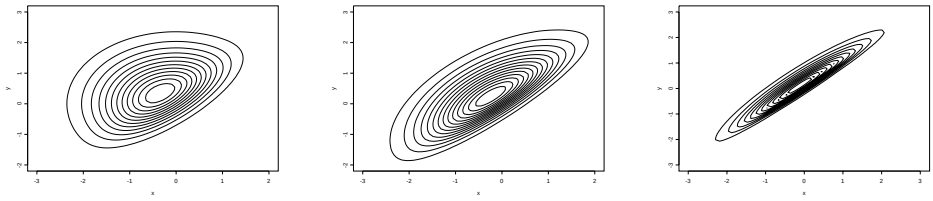


Figure 1: Contour of the bivariate skew-normal distribution in (2.2), with $\boldsymbol{\mu} = (0, 0)^T, \boldsymbol{\Sigma} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$ and $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)^T$ for different values of λ_1, λ_2 and ρ.

Proposition 1: Let $\mathbf{W} \sim SN_n(\boldsymbol{\lambda})$. Then

$$\mathbf{W} \stackrel{d}{=} \boldsymbol{\delta}|X_0| + (\mathbf{I}_n - \boldsymbol{\delta}\boldsymbol{\delta}^T)^{1/2}\mathbf{X}_1, \quad \textit{where} \quad \boldsymbol{\delta} = \frac{\boldsymbol{\lambda}}{\sqrt{1 + \boldsymbol{\lambda}^T\boldsymbol{\lambda}}}, \tag{2.5}$$

$X_0 \sim N_1(0, 1)$ independent of $\mathbf{X}_1 \sim N_n(\mathbf{0}, \mathbf{I}_n)$ and "$\stackrel{d}{=}$" meaning "distributed as".

In the appendix we provide a proof of this proposition. Notice that the stochastic representation give in Henze (1986) for the univariate case is a special case of (2.5). Thus, we have extended the univariate skew-normal distribution given in (2.1) in a nice way for the multivariate case. In Figure 1 we present some contours of the density associated with the bivariate skew-normal distribution $SN_2(\mathbf{0}, \boldsymbol{\Sigma}, \boldsymbol{\lambda})$ for different values of $\boldsymbol{\Sigma}$ and $\boldsymbol{\lambda}$. Note that these contours are not elliptical and can be strongly asymmetric depending on suitable choices of the parameters. A direct consequence of Proposition 1 is given in the following corollary.

Crollary 1: Let $\mathbf{Y} \overset{d}{=} \boldsymbol{\mu}+\boldsymbol{\Sigma}^{1/2}\mathbf{W}$, where $\mathbf{W} \sim SN_n(\boldsymbol{\lambda})$. Then $\mathbf{Y} \sim SN_n(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda})$. Moreover,

$$E[\mathbf{Y}] = \boldsymbol{\mu} + \sqrt{\frac{2}{\pi}}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\delta} \text{ and } Var[\mathbf{Y}] = \boldsymbol{\Sigma} - \frac{2}{\pi}\boldsymbol{\Sigma}^{1/2}\,\boldsymbol{\delta}\boldsymbol{\delta}^T\,\boldsymbol{\Sigma}^{1/2}.$$

2. A SNLMM Likelihood Function

In order to allow symmetric-asymmetric properties in characterizing features of real data sets, the SNLMM is defined by extending the normal mixed model in (1.1)-(1.2) by considering that

$$\mathbf{Y}_j = \mathbf{X}_j\boldsymbol{\beta} + \mathbf{Z}_j\mathbf{b}_j + \boldsymbol{\epsilon}_j, \tag{2.1}$$

$$\mathbf{b}_j \overset{\text{iid}}{\sim} SN_q(\mathbf{0}, \mathbf{D}, \boldsymbol{\lambda}_b),\ \ \boldsymbol{\epsilon}_j \overset{\text{ind}}{\sim} SN_{n_j}(\mathbf{0}, \boldsymbol{\psi}_j, \boldsymbol{\lambda}_{ej}),\ j = 1, \ldots, m\,, \tag{2.2}$$

with $\mathbf{b}_j$ independent of $\boldsymbol{\epsilon}_j$, which by using Corollary **??**, leads to the following hierarchical model:

$$\mathbf{Y}_j|\mathbf{b}_j \overset{\text{ind}}{\sim} SN_{n_j}(\mathbf{X}_j\boldsymbol{\beta} + \mathbf{Z}_j\mathbf{b}_j, \boldsymbol{\psi}_j, \boldsymbol{\lambda}_{ej}), \tag{2.3}$$

$$\mathbf{b}_j \overset{\text{iid}}{\sim} SN_q(\mathbf{0}, \mathbf{D}, \boldsymbol{\lambda}_b),\ \ j = 1, \ldots, m. \tag{2.4}$$

Note from (2.1)-(2.2) and Corollary 1 that,

$$E[\mathbf{Y}_j] = \mathbf{X}_j\boldsymbol{\beta} + E[\mathbf{Z}_j\mathbf{b}_j + \boldsymbol{\epsilon}_j] = \mathbf{X}_j\boldsymbol{\beta} + \sqrt{\frac{2}{\pi}}(\mathbf{Z}_j\mathbf{D}^{1/2}\boldsymbol{\delta}_b + \boldsymbol{\psi}_j^{1/2}\boldsymbol{\delta}_{ej}),$$

where $\boldsymbol{\delta}_b = \boldsymbol{\lambda}_b(1 + \boldsymbol{\lambda}_b^T\boldsymbol{\lambda}_b)^{-1/2}$ and $\boldsymbol{\delta}_{ej} = \boldsymbol{\lambda}_b(1 + \boldsymbol{\lambda}_{ej}^T\boldsymbol{\lambda}_{ej})^{-1/2}$, $j = 1, \ldots, m$, which must be considered in order to obtain a correct interpretation of the model parameters. However, in practice we can in general rescue the common interpretation, correcting the intercept parameter in the fitted model, as will be done in Section 4. The main interest is to make inference on the parameter vectors $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\alpha}^T, \boldsymbol{\gamma}^T)^T$ and $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_b^T, \boldsymbol{\lambda}_{e1}^T, \ldots, \boldsymbol{\lambda}_{em}^T)^T$. As discussed in Verbeke and Molenberghs (2000), unless the data are analyzed in a Bayesian framework, inference in this type of models has to be based on the marginal distribution for the response $\mathbf{Y}_j$. The marginal density of $\mathbf{Y}_j$ is obtained in the following theorem, the proof is given in the appendix.

Theorem 1: Let $\mathbf{Y}_j = \mathbf{X}_j\boldsymbol{\beta} + \mathbf{Z}_j\mathbf{b}_j + \boldsymbol{\epsilon}_j$, where $\mathbf{b}_j \overset{\text{ind}}{\sim} SN_q(\mathbf{0}, \mathbf{D}, \boldsymbol{\lambda}_b)$ and $\boldsymbol{\epsilon}_j \overset{\text{ind}}{\sim} SN_{n_j}(\mathbf{0}, \boldsymbol{\psi}_j, \boldsymbol{\lambda}_{e_j})$ are independent. Then, the marginal distribution of $\mathbf{Y}_j$ is given by

$$\begin{aligned} f_{\mathbf{Y}_j}(\mathbf{y}_j|\boldsymbol{\theta}, \boldsymbol{\lambda}) &= 2^2\phi_{n_j}(\mathbf{y}_j|\mathbf{X}_j\boldsymbol{\beta}, \boldsymbol{\Sigma}_j) \\ &\quad \times\Phi_2\left((\boldsymbol{\mu}_{2_j} - \boldsymbol{\Gamma}_j\boldsymbol{\mu}_{1_j})(\mathbf{y}_j - \mathbf{X}_j\boldsymbol{\beta})|\mathbf{0}, \mathbf{I}_2 + \boldsymbol{\Gamma}_j\boldsymbol{\Lambda}_j\boldsymbol{\Gamma}_j^T\right), \end{aligned} \tag{2.5}$$

where

$$\boldsymbol{\mu}_{1_j} = \boldsymbol{\Lambda}_j \mathbf{Z}_j^T \boldsymbol{\psi}_j^{-1}, \quad \boldsymbol{\Sigma}_j = \boldsymbol{\psi}_j + \mathbf{Z}_j \mathbf{D} \mathbf{Z}_j^T, \quad \boldsymbol{\Lambda}_j = (\mathbf{D}^{-1} + \mathbf{Z}_j^T \boldsymbol{\psi}_j^{-1} \mathbf{Z}_j)^{-1}, \tag{2.6}$$

$$\boldsymbol{\mu}_{2_j} = \begin{pmatrix} \boldsymbol{\lambda}_{e_j}^T \boldsymbol{\psi}_j^{-1/2} \\ \mathbf{0} \end{pmatrix} \quad \text{and} \quad \boldsymbol{\Gamma}_j = \begin{pmatrix} \boldsymbol{\lambda}_{e_j}^T \boldsymbol{\psi}^{-1/2} \mathbf{Z}_j \\ -\boldsymbol{\lambda}_b^T \mathbf{D}^{-1/2} \end{pmatrix}, \; j = 1, \ldots, m. \tag{2.7}$$

Note that the likelihood (2.5) is not in the class defined here, since the skewness factor in this expression is of dimension 2. It is, however, in the class of fundamental skew-normal distributions considered by Arellano-Valle and Genton (2005) (see also, (2.4) in Remark 1). The result presented in Theorem is important because it avoids using more complex numerical techniques such as Monte Carlo integration to carry out inference in this type of models, given that it allows a closed form for the marginal distribution of $\mathbf{Y}_j$, $j = 1, \ldots, m$, facilitating straightforward implementation of inferences with standard optimization routines. Thus, denoting the log-likelihood function by $\ell(\boldsymbol{\theta}, \boldsymbol{\lambda})$, it can be written as

$$\begin{aligned} \ell(\boldsymbol{\theta}, \boldsymbol{\lambda}) \;\propto\; & -\frac{1}{2} \sum_{j=1}^{m} \log |\boldsymbol{\Sigma}_j| - \frac{1}{2} \sum_{j=1}^{m} \{(\mathbf{y}_j - \mathbf{X}_j \boldsymbol{\beta})^T \boldsymbol{\Sigma}_j^{-1} (\mathbf{y}_j - \mathbf{X}_j \boldsymbol{\beta})\} \\ & + \sum_{j=1}^{m} \log \Phi_2 \left((\boldsymbol{\mu}_{2_j} - \boldsymbol{\Gamma}_j \boldsymbol{\mu}_{1_j})(\mathbf{y}_j - \mathbf{X}_j \boldsymbol{\beta}) | \mathbf{0}, \mathbf{I}_2 + \boldsymbol{\Gamma}_j \boldsymbol{\Lambda}_j \boldsymbol{\Gamma}_j^T \right), \end{aligned} \tag{2.8}$$

where $\boldsymbol{\mu}_{1j}$, $\boldsymbol{\mu}_{2j}$, $\boldsymbol{\Sigma}_j$, $\boldsymbol{\Gamma}_j$ and $\boldsymbol{\Lambda}_j$ as defined in (2.6) and (2.7).

We call attention to the fact that no explicit solution is available for the maximization problem so that the likelihood function has to be maximized numerically. Some special cases may be of interest. For instance, the situation where $\boldsymbol{\lambda}_{e1} = \ldots = \boldsymbol{\lambda}_{em} = \mathbf{0}$ or $\boldsymbol{\lambda}_b = \mathbf{0}$, which are special cases of the above general situation. These situations are treated next.

Corollary 2: Under the conditions of Theorem , it follows that:

(i) if $\boldsymbol{\lambda}_{e_j} = \mathbf{0}$, $j = 1, \ldots, m$, then

$$f_{\mathbf{Y}_j}(\mathbf{y}_j | \boldsymbol{\theta}, \boldsymbol{\lambda}_b) = 2 \phi_{n_j}(\mathbf{y}_j | \mathbf{X}_j \boldsymbol{\beta}, \boldsymbol{\Sigma}_j) \Phi_1 \left(\boldsymbol{\lambda}_{b_j}^{-T} \boldsymbol{\Sigma}_j^{-1/2} (\mathbf{y}_j - \mathbf{X}_j \boldsymbol{\beta}) \right), \tag{2.9}$$

i.e.,

$$\mathbf{Y}_j \sim SN_{n_j}(\mathbf{X}_j \boldsymbol{\beta}, \boldsymbol{\Sigma}_j, \bar{\boldsymbol{\lambda}}_{b_j}), \quad with \quad \bar{\boldsymbol{\lambda}}_{b_j} = \frac{\boldsymbol{\Sigma}_j^{-1/2} \mathbf{Z}_j \mathbf{D}^{1/2} \boldsymbol{\lambda}_b}{\sqrt{1 + \boldsymbol{\lambda}_b^T \mathbf{D}^{-1/2} \boldsymbol{\Lambda}_j \mathbf{D}^{-1/2} \boldsymbol{\lambda}_b}},$$

(ii) if $\boldsymbol{\lambda}_b = \mathbf{0}$, then

$$f_{\mathbf{Y}_j}(\mathbf{y}_j|\boldsymbol{\theta}, \boldsymbol{\lambda}_{e_j}) = 2\phi_{n_j}(\mathbf{y}_j|\mathbf{X}_j\boldsymbol{\beta}, \boldsymbol{\Sigma}_j)\Phi_1\left(\bar{\boldsymbol{\lambda}}_{e_j}^T\boldsymbol{\Sigma}_j^{-1/2}(\mathbf{y}_j - \mathbf{X}_j\boldsymbol{\beta})\right), \quad (2.10)$$

i.e.,

$$\mathbf{Y}_j \sim SN_{n_j}(\mathbf{X}_j\boldsymbol{\beta}, \boldsymbol{\Sigma}_j, \bar{\boldsymbol{\lambda}}_{e_j}), \quad with \quad \bar{\boldsymbol{\lambda}}_{e_j} = \frac{\boldsymbol{\Sigma}_j^{-1/2}\boldsymbol{\psi}_j^{1/2}\boldsymbol{\lambda}_{e_j}}{\sqrt{1 + \boldsymbol{\lambda}_{e_j}^T\boldsymbol{\psi}_j^{-1/2}\mathbf{Z}_j\boldsymbol{\Lambda}_j\mathbf{Z}_j^T\boldsymbol{\psi}_j^{-1/2}\boldsymbol{\lambda}_{e_j}}}.$$

Although simpler, the log-likelihood functions that follow from (2.9) and (2.10) must also be maximized numerically. The asymptotic covariance matrix of the maximum likelihood estimators (MLE) can be estimated by using the Hessian matrix, which can also be computed numerically by using the program R, for example. In the next section we present an EM-type algorithm for computing the MLE of densities obtained in Corollary 2.

4. An EM-type algorithm

A direct maximization of the likelihood (2.9) and (2.10) may sometimes poses problems since it involves terms like $\log(\Phi_1(w))$, which causes computational problems for negative w ($w < -40$, for example). Further, the approach seems not too robust with respect to starting values, that is, unless good starting values are used, the direct maximization approach will typically not converge. Simulation studies conducted indicate the EM to be more robust in the sense that it may converge more often than the direct maximization approach.

The EM algorithm (Dempster, Laird, and Rubin 1977) is a popular iterative algorithm for ML estimation in models with incomplete data. More specifically, let $\mathbf{y}$ denote the observed data and $\mathbf{t}$ denoted the missing data. The complete data $\mathbf{y}_{comp} = (\mathbf{y}, \mathbf{t})$ is $\mathbf{y}$ augmented with $\mathbf{t}$. We denote by $\ell_c(\boldsymbol{\theta}_c)$, $\boldsymbol{\theta}_c \in \boldsymbol{\Theta}$, the complete-data log-likelihood function and by $Q(\boldsymbol{\theta}_c|\boldsymbol{\theta}'_c)$ the expected complete-data log- likelihood

$$Q(\boldsymbol{\theta}_c|\boldsymbol{\theta}'_c) = E[\ell_c(\boldsymbol{\theta}_c)|\mathbf{y}, \boldsymbol{\theta}'_c]$$

Each iteration of the EM algorithm consist of two steps, the expectation step and the maximization step:

- E-step: Compute $Q(\boldsymbol{\theta}_c|\boldsymbol{\theta}_c^{(r)})$ as a function of $\boldsymbol{\theta}_c$;
- M-step: Find $\boldsymbol{\theta}_c^{(r+1)}$ such that $Q(\boldsymbol{\theta}_c^{(r+1)}|\boldsymbol{\theta}_c^{(r)}) = \max_{\boldsymbol{\theta}_c \in \boldsymbol{\Theta}} Q(\boldsymbol{\theta}_c|\boldsymbol{\theta}_c^{(r)})$.

Each iteration of the EM algorithm increases the likelihood function $\ell(\boldsymbol{\theta}_c)$ and the EM algorithm typically converges to a local or global maximum of the

likelihood function. When the M-step in the EM algorithm is difficult to implement, it is useful to replace it with a sequence of constrained maximization (CM) steps, each of which maximizes $Q(\boldsymbol{\theta}_c|\boldsymbol{\theta}_c^{(r)})$ over $\boldsymbol{\theta}_c$ with some function of $\boldsymbol{\theta}_c$ held fixed. The sequence of the CM-steps is such that the overall maximization is over the full parameter space. This leads to a simple extension of the EM algorithm, called the ECM algorithm (Meng and Rubin, 1993). In this work we implemented the ECM algorithm which irrespectively will be called EM-algorithm.

In order to implement the two steps of the EM-algorithm for maximizing the likelihood from Corollary , we need first some additional results. The proofs are given in the appendix.

Proposition 2: Suppose that $\mathbf{Y}|T = t \sim N_n(\boldsymbol{\mu} + \mathbf{d}t, \boldsymbol{\Psi})$ and $T \sim HN_1(0,1)$ (the standardized half-normal distribution). Let $\boldsymbol{\Sigma} = \boldsymbol{\Psi} + \mathbf{d}\mathbf{d}^T$. Then the joint distribution of $(\mathbf{Y}^T, T)^T$ can be written as

$$f_{\mathbf{Y},T}(\mathbf{y}, t|\boldsymbol{\theta}, \boldsymbol{\lambda}) = 2\phi_n(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma})\phi_1(t|\eta, \tau^2)\mathbb{I}\{t > 0\}, \tag{4.1}$$

where

$$\eta = \frac{\mathbf{d}^T\boldsymbol{\Psi}^{-1}(\mathbf{y} - \boldsymbol{\mu})}{1 + \mathbf{d}^T\boldsymbol{\Psi}^{-1}\mathbf{d}} \ \textit{and} \ \tau^2 = \frac{1}{1 + \mathbf{d}^T\boldsymbol{\Psi}^{-1}\mathbf{d}}. \tag{4.2}$$

Notice that the marginal distribution of $\mathbf{Y}$ follows from (4.1) after integrating out t and is given by

$$f_{\mathbf{Y}}(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\lambda}) = 2\phi_n(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma})\Phi_1\left(\frac{\eta}{\tau}\right). \tag{4.3}$$

Proposition 3: Under the conditions in Proposition ,

$$E[T^k|\mathbf{y}] = E[X^k|X > 0],$$

where $X \sim N_1(\eta, \tau^2)$, with η and τ^2 given in (4.2). Particularly,

$$E[T|\mathbf{y}] = \eta + \frac{\phi_1(\frac{\eta}{\tau})}{\Phi_1(\frac{\eta}{\tau})}\tau, \tag{4.4}$$

and

$$E(T^2|\mathbf{y}) = \eta^2 + \tau^2 + \frac{\phi_1(\frac{\eta}{\tau})}{\Phi_1(\frac{\eta}{\tau})}\tau\eta. \tag{4.5}$$

4.1 EM algorithm when $\boldsymbol{\lambda}_{e1} = \ldots = \boldsymbol{\lambda}_{em} = \mathbf{0}$

Under this assumption we have the following SNLMM:

$$\mathbf{Y}_j = \mathbf{X}_j\boldsymbol{\beta} + \mathbf{Z}_j\mathbf{b}_j + \boldsymbol{\epsilon}_j, \tag{4.6}$$

with

$$\boldsymbol{\epsilon}_j \overset{\text{ind}}{\sim} N_{n_j}(\mathbf{0}, \boldsymbol{\psi}_j), \quad \mathbf{b}_j \overset{\text{ind}}{\sim} SN_q(\mathbf{0}, \mathbf{D}, \boldsymbol{\lambda}_b), \ \ j = 1, \ldots, m. \tag{4.7}$$

It is clear that, (4.7) jointly with Proposition 1 implies that

$$\mathbf{b}_j \overset{d}{=} \mathbf{D}^{1/2}\boldsymbol{\delta}_b|X_{0j}| + \mathbf{D}^{1/2}(\mathbf{I}_q - \boldsymbol{\delta}_b\boldsymbol{\delta}_b^T)^{1/2}\mathbf{X}_{1j}, \ j = 1, \ldots, m, \tag{4.8}$$

where $X_{0j} \overset{\text{iid}}{\sim} N_1(0,1)$, $\mathbf{X}_{1j} \overset{\text{iid}}{\sim} N_q(\mathbf{0}, \mathbf{I}_q)$, with X_{0j} and $\mathbf{X}_{1j}$ independent $j = 1, \ldots, m$, and $\boldsymbol{\delta}_b = \dfrac{\boldsymbol{\lambda}_b}{\sqrt{1 + \boldsymbol{\lambda}_b^T\boldsymbol{\lambda}_b}}$. Moreover, independence between $\mathbf{b}_j$ and $\boldsymbol{\epsilon}_j$, $j = 1, \ldots, m$, imply that $\mathbf{V}_j = (X_{0j}, \mathbf{X}_{1j}^T)^T$ and $\boldsymbol{\epsilon}_j$, are independent, $j = 1, \ldots, m$. Hence, replacing (4.8) in (4.6) we have that

$$\mathbf{Y}_j = \mathbf{X}_j\boldsymbol{\beta} + \mathbf{Z}_j\bar{\boldsymbol{\delta}}_b t_j + \mathbf{r}_j, \tag{4.9}$$

where

$$\bar{\boldsymbol{\delta}}_b = \mathbf{D}^{1/2}\boldsymbol{\delta}_b, \quad t_j = |X_{0j}| \quad \text{and} \quad \mathbf{r}_j = \boldsymbol{\epsilon}_j + \mathbf{Z}_j\mathbf{D}^{1/2}(\mathbf{I}_q - \boldsymbol{\delta}_b\boldsymbol{\delta}_b^T)^{1/2}\mathbf{X}_{1j},$$

which are such that

$$\mathbf{r}_j \overset{\text{ind}}{\sim} N_{n_j}(\mathbf{0}, \boldsymbol{\psi}_j + \mathbf{Z}_j(\mathbf{D} - \bar{\boldsymbol{\delta}}_b\bar{\boldsymbol{\delta}}_b^T)\mathbf{Z}_j^T), \qquad t_j \overset{\text{iid}}{\sim} HN(0,1), \tag{4.10}$$

and are independent, $j = 1, \ldots, m$. Note that $\mathbf{r}_j$ has mean zero, so that it can be used, for example, in residual analysis to check model adequability. Besides, the second term on the right side of equation (4.9) has mean $\sqrt{\frac{2}{\pi}}\mathbf{Z}_j\bar{\boldsymbol{\delta}}_b$ which can be used to correct the model intercept so that the fixed effects have the same interpretation as in the usual LMM (population average).

Therefore, (4.9) and (4.10) imply that the model defined by (4.6)-(4.7) can be written as

$$\mathbf{Y}_j|t_j \overset{\text{ind}}{\sim} N_{n_j}(\boldsymbol{\mu}_j + \mathbf{d}_j t_j, \boldsymbol{\Psi}_j) \quad \text{and} \quad t_j \overset{\text{iid}}{\sim} HN_1(0,1), \ \ j = 1, \ldots, m, \tag{4.11}$$

where

$$\boldsymbol{\mu}_j = \mathbf{X}_j\boldsymbol{\beta}, \ \ \mathbf{d}_j = \mathbf{Z}_j\bar{\boldsymbol{\delta}}_b, \ \ \boldsymbol{\Psi}_j = \boldsymbol{\Sigma}_j - \mathbf{d}_j\mathbf{d}_j^T \ \text{ and } \ \boldsymbol{\Sigma}_j = \boldsymbol{\psi}_j + \mathbf{Z}_j\mathbf{D}\mathbf{Z}_j^T. \tag{4.12}$$

Note that in (4.12) $\boldsymbol{\mu}_j$ and $\boldsymbol{\Sigma}_j$ are the marginal mean vector and covariance matrix, respectively, under the usual linear mixed model. Let $\mathbf{y}^T = (\mathbf{y}_1^T, \ldots, \mathbf{y}_m^T)^T$ and $\mathbf{t} = (t_1, \ldots, t_m)^T$, as a direct consequence of Proposition using simple algebra we have the next result.

Proposition 4: Under (4.11) it follows that the complete log-likelihood function associated with $(\mathbf{y}, \mathbf{t})$ in the SNLMM (4.6)-(4.7), can be written as

$$\ell_c(\boldsymbol{\theta}, \boldsymbol{\lambda}_b) \propto -\frac{1}{2}\sum_{j=1}^{m}\log|\boldsymbol{\Psi}_j| - \frac{1}{2}\sum_{j=1}^{m}(\mathbf{y}_j - \boldsymbol{\mu}_j)^T\boldsymbol{\Sigma}_j^{-1}(\mathbf{y}_j - \boldsymbol{\mu}_j) - \frac{1}{2}\sum_{j=1}^{m}\frac{(t_j - \eta_j)^2}{\tau_j^2}, \tag{4.13}$$

where by (4.2)

$$\eta_j = \frac{\mathbf{d}_j^T\boldsymbol{\Psi}_j^{-1}(\mathbf{y}_j - \boldsymbol{\mu}_j)}{1 + \mathbf{d}_j^T\boldsymbol{\Psi}_j^{-1}\mathbf{d}_j} \;\; and \;\; \tau_j^2 = \frac{1}{1 + \mathbf{d}_j^T\boldsymbol{\Psi}_j^{-1}\mathbf{d}_j}, \tag{4.14}$$

with $\boldsymbol{\mu}_j$, $\mathbf{d}_j$, $\boldsymbol{\Sigma}_j$ and $\boldsymbol{\Psi}_j$ as defined in (4.12).

Letting $\boldsymbol{\theta}_c = (\boldsymbol{\theta}^T, \boldsymbol{\lambda}_b^T)^T$, $\widehat{t}_j = E(T_j|\widehat{\boldsymbol{\theta}}_c, \mathbf{Y}_j = \mathbf{y}_j)$ and $\widehat{t_j^2} = E(T_j^2|\widehat{\boldsymbol{\theta}}_c, \mathbf{Y}_j = \mathbf{y}_j)$, we obtain from Preposition **??** that

$$\widehat{t}_j = \widehat{\eta}_j + \frac{\phi_1(\frac{\widehat{\eta}_j}{\widehat{\tau}_j})}{\Phi_1(\frac{\widehat{\eta}_j}{\widehat{\tau}_j})}\widehat{\tau}_j, \tag{4.15}$$

$$\widehat{t_j^2} = \widehat{\eta_j^2} + \widehat{\tau_j^2} + \frac{\phi_1(\frac{\widehat{\eta}_j}{\widehat{\tau}_j})}{\Phi_1(\frac{\widehat{\eta}_j}{\widehat{\tau}_j})}\widehat{\tau}_j\widehat{\eta}_j, \tag{4.16}$$

where η_j and τ_j^2 as in (4.14). We then have the following EM algorithm:

E-step: Given $\boldsymbol{\theta}_c = \widehat{\boldsymbol{\theta}}_c$, compute $\widehat{t}_j$ and $\widehat{t_j^2}$ for $j = 1, \ldots, m$, using (4.15) and (4.16), respectively.

M-step: Update $\widehat{\boldsymbol{\theta}}_c$ by maximizing $E[\ell_c(\boldsymbol{\theta}_c)|\mathbf{y}, \widehat{\boldsymbol{\theta}}_c]$ over $\boldsymbol{\theta}_c$, which leads to

$$\begin{aligned}\widehat{\boldsymbol{\beta}} &= [\sum_{j=1}^{m}\mathbf{X}_j^T(\widehat{\boldsymbol{\Sigma}}_j^{-1} + \widehat{\tau}_j^2\widehat{\boldsymbol{\Psi}}_j^{-1}\widehat{\mathbf{d}}_j\widehat{\mathbf{d}}_j^T\widehat{\boldsymbol{\Psi}}_j^{-1})\mathbf{X}_j]^{-1} \times \\ &\quad \sum_{j=1}^{m}[\mathbf{X}_j^T(\widehat{\boldsymbol{\Sigma}}_j^{-1} + \widehat{\tau}_j^2\widehat{\boldsymbol{\Psi}}_j^{-1}\widehat{\mathbf{d}}_j\widehat{\mathbf{d}}_j^T\widehat{\boldsymbol{\Psi}}_j^{-1})\mathbf{y}_j - \widehat{t}_j\mathbf{X}_j^T\widehat{\boldsymbol{\Psi}}_j^{-1}\widehat{\mathbf{d}}_j],\end{aligned} \tag{4.17}$$

and

$$\hat{\boldsymbol{\nu}} = \text{argmax}_{\boldsymbol{\nu}}[\ell_c(\widehat{\boldsymbol{\beta}}, \boldsymbol{\nu})], \;\; \text{with} \;\; \boldsymbol{\nu} = (\boldsymbol{\alpha}^T, \boldsymbol{\gamma}^T, \boldsymbol{\lambda}_b^T)^T, \tag{4.18}$$

where $\ell_c(\widehat{\boldsymbol{\beta}}, \boldsymbol{\nu})$ is (4.13) evaluated at updated $\widehat{\boldsymbol{\beta}}$, $t_j = \widehat{t}_j$ and $t_j^2 = \widehat{t_j^2}$, $j = 1, \ldots, m$.

4.2 EM algorithm when $\boldsymbol{\lambda}_b = \mathbf{0}$

Likewise, considering the case where $\boldsymbol{\lambda}_b = \mathbf{0}$, that is, the linear mixed model in (4.6), with the assumption that

$$\boldsymbol{\epsilon}_j \overset{\text{ind}}{\sim} SN_{n_j}(\mathbf{0}, \boldsymbol{\psi}_j, \boldsymbol{\lambda}_{ej}) \quad \text{and} \quad \mathbf{b}_j \overset{\text{iid}}{\sim} N_q(\mathbf{0}, \mathbf{D}), \quad j = 1, \ldots, m, \tag{4.19}$$

all independent, we can write

$$\mathbf{Y}_j = \mathbf{X}_j\boldsymbol{\beta} + \boldsymbol{\psi}_j^{1/2}\boldsymbol{\delta}_{ej}t_j + \mathbf{r}_j, \tag{4.20}$$

where

$$t_j = |X_{1j}|, \;\; \boldsymbol{\delta}_{ej} = \frac{\boldsymbol{\lambda}_{ej}}{(1+\boldsymbol{\lambda}_{ej}\boldsymbol{\lambda}_{ej}^T)^{1/2}} \;\text{ and }\; \mathbf{r}_j = \mathbf{Z}_j\mathbf{b}_j + \boldsymbol{\psi}_j^{1/2}(\mathbf{I}_{n_j} - \boldsymbol{\delta}_{ej}\boldsymbol{\delta}_{ej}^T)^{1/2}\mathbf{X}_{0j},$$

this is

$$\mathbf{Y}_j|t_j \overset{\text{ind}}{\sim} N_{n_j}(\boldsymbol{\mu}_j + \mathbf{d}_j t_j, \boldsymbol{\Psi}_j) \;\text{ and }\; t_j \sim HN_1(0,1), \;\; j = 1, \ldots, m, \tag{4.21}$$

where

$$\boldsymbol{\mu}_j = \mathbf{X}_j\boldsymbol{\beta}, \;\; \mathbf{d}_j = \boldsymbol{\psi}_j^{1/2}\boldsymbol{\delta}_{ej}, \;\; \boldsymbol{\Psi}_j = \boldsymbol{\Sigma}_j - \mathbf{d}_j\mathbf{d}_j^T \;\text{ and }\; \boldsymbol{\Sigma}_j = \boldsymbol{\psi}_j + \mathbf{Z}_j\mathbf{D}\mathbf{Z}_j^T. \tag{4.22}$$

As a consequence of the above results, it follows from Proposition 2 that:

Proposition 5: Under (4.21) it follows that the complete log-likelihood function associated with $(\mathbf{y}, \mathbf{t})$ in the SNLMM (4.6) and (4.19), can be written as

$$\ell_c(\boldsymbol{\theta}, \boldsymbol{\lambda}_e) \propto -\frac{1}{2}\sum_{j=1}^{m}\log|\boldsymbol{\Psi}_j| - \frac{1}{2}\sum_{j=1}^{m}(\mathbf{y}_j - \boldsymbol{\mu}_j)^T\boldsymbol{\Sigma}_j^{-1}(\mathbf{y}_j - \boldsymbol{\mu}_j) - \frac{1}{2}\sum_{j=1}^{m}\frac{(t_j - \eta_j)^2}{\tau_j^2}, \tag{4.23}$$

where

$$\eta_j = \frac{\mathbf{d}_j^T\boldsymbol{\Psi}_j^{-1}(\mathbf{y}_j - \boldsymbol{\mu}_j)}{1 + \mathbf{d}_j^T\boldsymbol{\Psi}_j^{-1}\mathbf{d}_j} \;\text{ and }\; \tau_j^2 = \frac{1}{1 + \mathbf{d}_j^T\boldsymbol{\Psi}_j^{-1}\mathbf{d}_j}, \tag{4.24}$$

with $\boldsymbol{\mu}_j$, $\mathbf{d}_j$, $\boldsymbol{\Sigma}_j$ and $\boldsymbol{\Psi}_j$ defined as in (4.22).

The EM algorithm in this case proceed as in (4.15)-(4.18), with η_j and τ_j^2 as in (4.24). Note that in both cases the M-step require numerical maximization which can be easily implemented using statistical software as Ox, R and Matlab with bfgs, optim and fmincon routines, respectively. The starting values are often chosen to be the corresponding estimates under a normal assumption, where the starting values for the asymmetric parameters are set to be $\mathbf{0}$ and as recommended

in the literature, it is useful to run the EM-algorithm several times with different starting values.

Following other authors (e.g. Zhang and Davidian, 2001) we propose evaluate a series of fits by inspection of information criteria such as Akaike's Information Criterion (AIC, $-\ell(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\lambda}})/N + P/N$), Schawarz's Bayesian Information Criterion (BIC, $-\ell(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\lambda}})/N + 0.5\log(N)P/N$), and the Hannan-Quinn Criterion (HQ, $-\ell(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\lambda}})/N + \log(\log(N))P/N$), where P is the number of free parameters in the model and $N = \sum_{j=1}^{m} n_j$, as we demonstrate shortly in the next section.

Table 1: Monte Carlo results based on 100 data sets, true $Gamma(4, 1)$ distribution for the random effects. MC Mean and MC SD are average and standard deviation of the estimates, AVE SE is average of estimated standard errors. EC is the empirical coverage probability of the 95% confidence intervals of the estimates. True values of parameters are in parentheses.

Parameter	MC Mean	MC SD	AVE SE	EC
(a) Skew-Normal Scenario				
α	1.5447	0.3349	-	-
$\beta_1(2)$	2.0008	0.0106	0.0108	0.95
$\beta_2(1)$	0.9731	0.2081	0.2083	0.95
$\sigma_e^2(0.25)$	0.2490	0.0135	0.0119	0.93
$E[\alpha + b]$ (4)	4.0083	0.1636	-	-
$Var[\alpha + b](4)$	3.9853	0.5399	-	-
(b) Normal Scenario				
α	4.0213	0.1828	-	-
$\beta_1(2)$	2.0008	0.0106	0.0116	0.98
$\beta_2(1)$	0.9336	0.2638	0.3035	0.93
$\sigma_e^2(0.25)$	0.2490	0.0135	0.0125	0.93
$E[\alpha + b]$ (4)	4.0213	0.1828	-	-
$Var[\alpha + b](4)$	3.9409	0.5839	-	-

5. Simulation Study

To assess the performance of the proposed model and methods, we conducted two simulation studies. In all cases, we took the linear mixed model to be

$$Y_{ij} = \alpha + t_{ij}\beta_1 + w_j\beta_2 + b_j + e_{ij}, \tag{5.1}$$

where for $i = 1, \ldots, 5$, $t_{ij} = i - 3$, $\beta_1 = 2$, $\beta_2 = 1$, and $e_{ij} \sim N(0, 0.5^2)$. In each simulation, 100 Monte Carlo data sets were simulated from (5.1) according to the

additional specifications described below. As an advantage of the skew-normal representation for the random effects is its propensity for accommodating skewness, in the first simulation we generated the $\alpha + b_j$ according to a $Gamma(4, 1)$ distribution with probability density $(1/6)x^3 \exp(-x)$, yielding a highly skewed distribution as suggested by Figure 2. Under this specification, $E[\alpha + b_j] = 4$, and $Var[\alpha + b_j] = 4$. Note that t_{ij} represents a covariate with values changing within individuals and the same for all individuals, while w_j is the individual level-covariate, e.g., a treatment indicator. We took $m = 200$ with $w_j = 1$ if $j \leq 100$ and $w_j = 0$ if $j > 100$. For each of 100 Monte Carlo generated data sets, (5.1) was fit two times under the assumption of previous section, with the density of b_j represented by the skew-normal distribution and also by the normal distribution. For evaluating the objective use of the criteria, the model preferred by each of AIC, BIC and HQ was recorded.

Table 2: Monte Carlo results based on 100 data sets, true $Normal(0, 4)$ distribution for the random effects. Entries are as in Table 1.

Parameter	MC Mean	MC SD	AVE SE	EC
(a) Skew-Normal Scenario				
$\alpha\,(4)$	3.6766	1.1681	1.3058	0.60
$\beta_1\,(2)$	2.0009	0.0103	0.0108	0.96
$\beta_2\,(1)$	1.0161	0.3137	0.2953	0.92
$\sigma_e^2(0.25)$	0.2482	0.0143	0.0121	0.89
$E[\alpha + b]\ (4)$	3.9799	0.1916	-	-
$Var[\alpha + b](4)$	3.9368	0.4306	-	-
(b) Normal Scenario				
$\alpha\,(4)$	3.9780	0.1923	0.2056	0.96
$\beta_1\,(2)$	2.0009	0.0103	0.0111	0.96
$\beta_2\,(1)$	1.0207	0.3167	0.2932	0.95
$\sigma_e^2\,(0.25)$	0.2482	0.0143	0.0124	0.90
$E[\alpha + b]\ (4)$	3.9780	0.1923	-	-
$Var[\alpha + b](4)$	3.9347	0.4312	-	-

6. An Application

We illustrate the usefulness of the proposed methods by applying them to longitudinal data on cholesterol levels collected as part of the famed Framingham heart study. The file includes the cholesterol levels over time, age at baseline and gender for $m = 200$ randomly selected individuals, reported in Zhang and

Davidian (2001). We adopt the same linear mixed model used by these authors, given by

$$y_{ij} = \beta_o + \beta_1 sex_j + \beta_2 age_j + \beta_3 t_{ij} + b_{oj} + b_{1j} t_{ij} + \epsilon_{ij}\,, \tag{6.1}$$

where y_{ij} is cholesterol level divided by 100 at the i-th time for subject j and t_{ij} is $(time - 5)/10$, with time measured in years from baseline; age_j is age at baseline; sex_j is the gender indicator ($0 = female$, $1 = male$). Thus, $\mathbf{x}_{ij} = (1,\, age_j,\, sex_j, t_{ij})^T$, $\mathbf{b}_j = (b_{oj},\, b_{1j})^T$ and $\mathbf{Z}_j = (1,\, t_{ij})^T$. Figure 3(a) shows the histogram of cholesterol levels, clearly indicating its asymmetric nature and that it seems adequate fitting a skew-normal model to the data set. Zhang and Davidian (2001) analyzed this data and show that the asymmetric behavior is partially explained by the available covariates and the random effects may not be normally distributed. Based in this information, three statistical models, differing in the error term and random effects distributions, are entertained. These models are:

Model 1: A model with independent multivariate normal distribution for the errors and multivariate skew-normal distribution for random effects with $\boldsymbol{\lambda}_b = (\lambda_{b1},\, \lambda_{b2})^T$;

Model 2: A model with independent multivariate skew -normal distribution for random random errors with common shape parameter between groups and multivariate symmetric normal distribution for the random effects; and

Model 3: A purely Gaussian model.

For the EM algorithm, none of AIC, BIC, or HQ selected the normal specification for any of the 100 data sets, demonstrating the ability of these selection methods to detect an obvious departure from normality and suggesting strong evidence of skewness. Table **??** shows the numerical results for the estimates where normality and skew-normal was considered. For the most part, parameter estimates are unbiased. In the most cases, the average of estimates standard errors agrees well with the Monte Carlo standard deviation. As found by other authors (e.g., Tao et al., 1999; Zhang and Davidian, 2001), efficiency of estimation on β_2 associated with the individual-level covariate w_j is degraded when normality is assumed relative to allowing a more flexible representation via the skew-normal distribution. Because one of the main focuses of such analysis may well be evaluation of treatment effect, this suggests that adopting the normality assumptions routinely may lead to inefficient inferences on fixed effects of primary interest. Note that some efficiency loss is also associated with estimation of the inter-individual variance $Var[b_j]$. Thus, although unbiased estimation is still possible under normality, failure to take appropriate account of the true features of the random effects leads to less precise inference on what are usually quantities of key interest.

The advantage of estimating the random effects density may be appreciated from Figure 2. Figure 2(a) shows the Monte Carlo average of density estimates over the 100 data sets along with the true density, the symmetric-normal fit and the fit for the skew-normal. The figure demonstrates that the additional flexibility afforded by the skew-normal representation is sufficient to capture quite accurately the true underlying features of the random effects. This observation is further supported by Figure 2(b), which shows the 100 density estimates from the skew-normal fits.

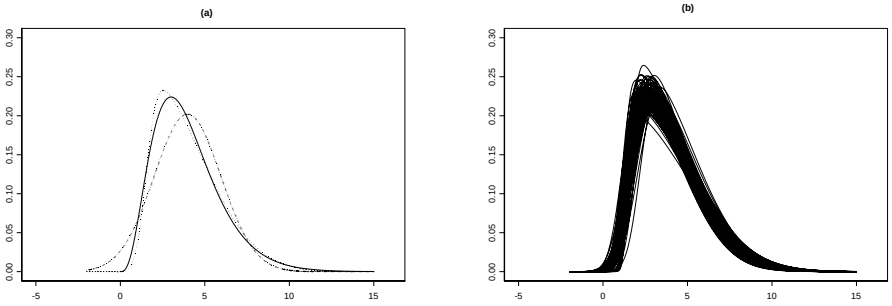


Figure 2: Simulation results based on 100 data sets. (a) True density (solid line) and Monte Carlo average estimated densities for 100 data set: using normal (dashed-dotted) and using skew-normal (dotted-line) (b) Estimated densities for the skew-normal fits.

The second simulation was identical to the first except that the true distribution of $\alpha + b_j$ was instead $N(4, 4)$, that is, $\alpha = 4$ and $b_j \sim N(0, 4)$. Here, then, there is no need for greater flexibility on the random effect, and the hope would be that the proposed methods would identify this. The AIC, BIC and HQ criteria correctly selected the normal distribution 86%, 98%, and 93% of the time, respectively. Summaries of the Monte Carlo result are given in Table 2, which shows that there is no efficiency loss associated with using the skew-normal distribution. The apparent conclusion is that the price to pay for estimating the random effects density when the normality assumption holds is mild; similar results are reported in Hu, Tsiatis, and Davidian (1998) and Zhang and Davidian (2001).

In all cases we considered $\boldsymbol{\psi}_j = \sigma_e^2 \mathbf{I}_{n_j}, j = 1, \ldots, 200$ (conditional independence). Table 3 presents the results obtained using the EM-type algorithm of the three models described above, SE are the estimated asymptotic standard errors based in the Hessian matrix computed numerically. When considering **Model 2** (only random errors are asymmetrically distributed), asymmetry is not detected and parameter estimates are close to the ones obtained under normality (**Model 3**). The AIC, BIC and HQ criteria indicate that **Model 1** presents the best

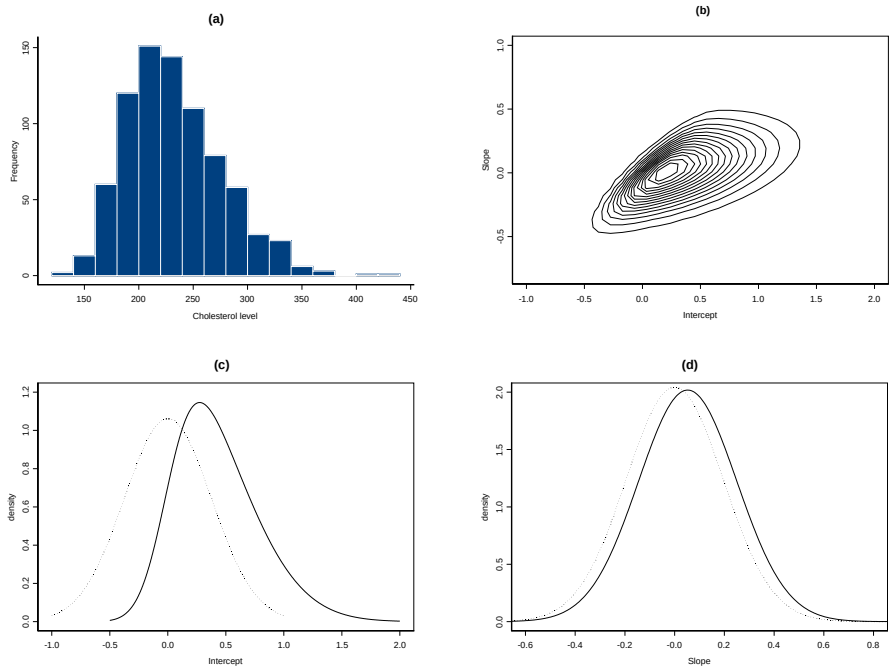
fit, supporting the contention of a departure from normality. Estimates of the individual-level covariate coefficients β_1, β_2 and β_3 differ as well, reflecting the qualitative behavior seen in the simulations. Figure 3(b)-(d) shows the estimated random effects distribution which is obvious skewness that is particularly evident in the contour plot showed in Figure 3(b). Figures 3(c) and (d) show the estimated marginal densities of b_{oj} and b_{1j}, respectively. Note that the distribution of slopes appears normal, while the shape of the density for intercepts shows evidence of skewness as Zhang and Davidian (2001) concluded using the SNP representation.

Table 3: Results of fitting models 1, 2 and 3 to the cholesterol data. d_{11}, d_{12} and d_{22} are the distinct elements of the matrix $\mathbf{D}^{1/2}$

	Model 1		Model 2		Model 3	
Parameter	Estimate	SE	Estimate	SE	Estimate	SE
β_o	1.3755	0.1418	1.5955	0.1568	1.5968	0.1543
β_1	−0.0591	0.0534	−0.0630	0.0554	−0.0630	0.0568
β_2	0.0136	0.0034	0.0184	0.0035	0.0184	0.0037
β_3	0.2281	0.0511	0.2817	0.0240	0.2817	0.0242
σ_e^2	0.0434	0.0025	0.0434	0.0024	0.0434	0.0024
d_{11}	0.5600	0.0418	0.3716	0.0199	0.3715	0.0201
d_{12}	0.0700	0.0317	0.0563	0.0173	0.0563	0.0179
d_{22}	0.1924	0.0311	0.1868	0.0308	0.1868	0.0329
λ_{b1}	2.9947	0.7789	-	-	-	-
λ_{b2}	0.0000	0.4814	-	-	-	-
λ_e	-	-	0.0140	0.9890	-	-
−log-likelihood	−666.1717		−659.7560		−659.7560	
AIC	−0.6285		−0.6233		−0.6243	
BIC	−0.6057		−0.6020		−0.6053	
HQ	−0.6195		−0.6152		−0.6171	

7. Final Conclusion

In this paper we developed a skew-normal mixed models for fitting regression model with dependent data. We believe that this is the first attempt in working in such general distributional structure for mixed models and that the approach used in this paper can be used in treating other multivariate models which will be the subject of incoming papers. An analytical expression (closed form) is



Figuree 3: (a) Histogram of cholesterol levels for 200 subjects of Framingham cholesterol study. (b) Contour plot of estimated density of $\mathbf{b}_j$. (c) and (d) Corresponding estimated marginal densities for components of $\mathbf{b}_j$ (solid) with estimated normal density (dotted) superimposed.

obtained for the marginal distribution of the observed response vector which allows carrying out inferences using standard optimization techniques and existing statistical softwares. For evaluation of the MLE, an EM-type algorithm is developed by exploring statistical properties of the model considered, and as is typical for the EM algorithm, reliability rather than speed is its best feature. An small simulation study is also presented where as observed in other contexts and approaches (e.g., Zhang and Davidian, 2001), there is potencial to gain efficiency in estimating certain parameters when the normality assumption does not hold, with only a small price to pay for the extra complications in the assumptions. An additional major advantage of all approaches that relax the assumptions on the random effects and the model errors densities is the insight the estimates provide. We have implemented the approach using *MATLAB* with the *fmincon* optimizer routine; code is available from the authors upon request.

Acknowledgments

The first author was partially supported by grant Fondecyt-1040865, Chile. Grants from CNPq/CAPES – Brazil are also acknowledged.

Appendix: Proofs

Before proving the results we consider the following lemmas. The notation used is that of Section 2.

Lemma 1. Let $\mathbf{Y} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then for any fixed k-dimensional vector $\mathbf{a}$ and $k \times n$ matrix $\mathbf{B}$,

$$E[\Phi_k(\mathbf{a} + \mathbf{B}\mathbf{Y}|\boldsymbol{\eta}, \boldsymbol{\Omega})] = \Phi_k(\mathbf{a}|\boldsymbol{\eta} - \mathbf{B}\boldsymbol{\mu}, \boldsymbol{\Omega} + \mathbf{B}\boldsymbol{\Sigma}\mathbf{B}^T).$$

Proof. The proof follows by noticing that

$$E[\Phi_k(\mathbf{a} + \mathbf{B}\mathbf{Y}|\boldsymbol{\eta}, \boldsymbol{\Omega})] = E[P(\mathbf{U} \leq \mathbf{a}|\mathbf{Y})] = P(\mathbf{U} \leq \mathbf{a}),$$

where $\mathbf{U}|\mathbf{Y} = \mathbf{y} \sim N_k(\boldsymbol{\eta} - \mathbf{B}\mathbf{y}, \boldsymbol{\Omega})$, so that $\mathbf{U} \sim N_k(\boldsymbol{\eta} - \mathbf{B}\boldsymbol{\mu}, \boldsymbol{\Omega} + \mathbf{B}\boldsymbol{\Sigma}\mathbf{B}^T)$.

Lemma 2. Let $\mathbf{Y} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\mathbf{X} \sim N_q(\boldsymbol{\eta}, \boldsymbol{\Omega})$. Then,

$$\begin{aligned} \phi_p(\mathbf{y}|\boldsymbol{\mu} + \mathbf{A}\mathbf{x}, \boldsymbol{\Sigma})\phi_q(\mathbf{x}|\boldsymbol{\eta}, \boldsymbol{\Omega}) &= \phi_p(\mathbf{y}|\boldsymbol{\mu} + \mathbf{A}\boldsymbol{\eta}, \boldsymbol{\Sigma} + \mathbf{A}\boldsymbol{\Omega}\mathbf{A}^T) \\ &\times \phi_q(\mathbf{x}|\boldsymbol{\eta} + \boldsymbol{\Lambda}\mathbf{A}^T\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \boldsymbol{\mu} - \mathbf{A}\boldsymbol{\eta}), \boldsymbol{\Lambda}), \end{aligned}$$

where $\boldsymbol{\Lambda} = (\boldsymbol{\Omega}^{-1} + \mathbf{A}^T\boldsymbol{\Sigma}^{-1}\mathbf{A})^{-1}$.

Proof. By letting $\mathbf{z} = \mathbf{y} - \boldsymbol{\mu} - \mathbf{A}\boldsymbol{\eta}$ and $\mathbf{w} = \mathbf{x} - \boldsymbol{\eta}$, we have after some standard algebraic operations that

$$\begin{aligned} (\mathbf{z} - \mathbf{A}\mathbf{w})^T\boldsymbol{\Sigma}^{-1}(\mathbf{z} - \mathbf{A}\mathbf{w}) &+ \mathbf{w}^T\boldsymbol{\Omega}^{-1}\mathbf{w} = \\ \mathbf{z}(\boldsymbol{\Sigma} + \mathbf{A}\boldsymbol{\Omega}\mathbf{A}^T)^{-1}\mathbf{z} &+ (\mathbf{w} - \boldsymbol{\Lambda}\mathbf{A}^T\boldsymbol{\Sigma}^{-1}\mathbf{z})^T\boldsymbol{\Lambda}^{-1}(\mathbf{w} - \boldsymbol{\Lambda}\mathbf{A}^T\boldsymbol{\Sigma}^{-1}\mathbf{z}), \end{aligned}$$

and the proof follows by noting also that $|\boldsymbol{\Sigma} + \mathbf{A}\Omega\mathbf{A}^T||\boldsymbol{\Lambda}| = |\boldsymbol{\Sigma}||\boldsymbol{\Omega}|$.

Proof of Preposition 1:

Let $\mathbf{U} = \boldsymbol{\delta}|X_0| + (\mathbf{I}_n - \boldsymbol{\delta}\boldsymbol{\delta}^T)^{1/2}\mathbf{X}_1$. Since $\mathbf{U}||X_0| = t \sim N_n(\boldsymbol{\delta}t, \mathbf{I}_n - \boldsymbol{\delta}\boldsymbol{\delta}^T)$, where $|X_0| \sim HN(0, 1)$, then by Lemma **??** it follows that

$$\begin{aligned} f_{\mathbf{U}}(\mathbf{w}) &= \int_0^\infty \phi_n(\mathbf{w}|\boldsymbol{\delta}t, \mathbf{I}_n - \boldsymbol{\delta}\boldsymbol{\delta}^T)2\phi(t)dt = \int_0^\infty \phi_n(\mathbf{w}|\mathbf{0}, \mathbf{I}_n)2\phi(t|\boldsymbol{\delta}^T\mathbf{w}, 1 - \boldsymbol{\delta}^T\boldsymbol{\delta})dt \\ &= 2\phi_n(\mathbf{w})\Phi_1\left(\frac{\boldsymbol{\delta}^T\mathbf{w}}{\sqrt{1 - \boldsymbol{\delta}^T\boldsymbol{\delta}}}\right), \end{aligned}$$

i.e., $\mathbf{U} \stackrel{d}{=} \mathbf{W} \sim SN_n(\boldsymbol{\lambda})$, with $\boldsymbol{\lambda} = \frac{\boldsymbol{\delta}}{\sqrt{1-\boldsymbol{\delta}^T\boldsymbol{\delta}}}$, which concludes the proof.

Proof of Theorem 1:

To obtain the marginal distribution of $\mathbf{Y}_j$, we drop the subscript j. From (2.3), (2.4) and the definition in (2.2), it follows that the marginal density of $\mathbf{Y}$ is obtained by computing the following integral:

$$\begin{aligned} f_{\mathbf{Y}}(\mathbf{y}|\boldsymbol{\theta},\boldsymbol{\lambda}) &= \int_{\mathbb{R}^q} f(\mathbf{y}|\mathbf{b},\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\lambda}_e) f(\mathbf{b}|\boldsymbol{\alpha},\boldsymbol{\lambda}_b) d\mathbf{b} \\ = \int_{\mathbb{R}^q} & 2^2 \phi_n(\mathbf{y}|\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\mathbf{b},\boldsymbol{\psi}) \Phi_1(\boldsymbol{\lambda}_e^T \boldsymbol{\psi}^{-1/2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{b})) \\ & \times \phi_q(\mathbf{b}|\mathbf{0},\mathbf{D}) \Phi_1(\boldsymbol{\lambda}_b^T \mathbf{D}^{-1/2}\mathbf{b}) d\mathbf{b}, \end{aligned} \tag{A.1}$$

which is based on the following lemma:

Lemma 3. Under the notation considered above, it follows that

$$\phi_n(\mathbf{y}|\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\mathbf{b},\boldsymbol{\psi})\phi_q(\mathbf{b}|\mathbf{0},\mathbf{D}) = \phi_n(\mathbf{y}|\mathbf{X}\boldsymbol{\beta},\boldsymbol{\Sigma})\phi_q(\mathbf{b}|\boldsymbol{\mu}_1(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}),\boldsymbol{\Lambda}) \tag{A.2}$$

and

$$\Phi_1(\boldsymbol{\lambda}_e^T \boldsymbol{\psi}^{-1/2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{b}))\Phi_1(\boldsymbol{\lambda}_b^T \mathbf{D}^{-1/2}\mathbf{b}) = \Phi_2(-\boldsymbol{\Gamma}\mathbf{b}|-\boldsymbol{\mu}_2(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}),\mathbf{I}_2), \tag{A.3}$$

where

$$\boldsymbol{\mu}_1 = \boldsymbol{\Lambda}\mathbf{Z}^T\boldsymbol{\psi}^{-1}, \quad \boldsymbol{\Sigma} = \boldsymbol{\psi}+\mathbf{Z}\mathbf{D}\mathbf{Z}^T, \quad \boldsymbol{\Lambda} = (\mathbf{D}^{-1}+\mathbf{Z}^T\boldsymbol{\psi}^{-1}\mathbf{Z})^{-1}, \tag{A.4}$$

$$\boldsymbol{\mu}_2 = \begin{pmatrix} \boldsymbol{\lambda}_e^T \boldsymbol{\psi}^{-1/2} \\ \mathbf{0} \end{pmatrix} \quad \text{and} \quad \boldsymbol{\Gamma} = \begin{pmatrix} \boldsymbol{\lambda}_e^T \boldsymbol{\psi}^{-1/2}\mathbf{Z} \\ -\boldsymbol{\lambda}_b^T \mathbf{D}^{-1/2} \end{pmatrix}. \tag{A.5}$$

Proof. Result (A.2) follows from Lemma **??**. Result (A.3) is proved by noting that if U and V are i.i.d. $N(0,1)$ random variables, then (A.3) can be written as

$$P(U \le \boldsymbol{\lambda}_e^T \boldsymbol{\psi}^{-1/2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{b}))P(V \le \boldsymbol{\lambda}_b^T \mathbf{D}^{-1/2}\mathbf{b}) = P(\mathbf{T} \le \boldsymbol{\mu}_2(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})),$$

where $\mathbf{T} = \mathbf{W}+\boldsymbol{\Gamma}\mathbf{b}$, with $\mathbf{W} = (U,V)^T \sim N_2(\mathbf{0},\mathbf{I}_2)$, and $\boldsymbol{\mu}_2$, $\boldsymbol{\Gamma}$ defined as in (A.5). Thus, since $\mathbf{T} \sim N_2(\boldsymbol{\Gamma}\mathbf{b},\mathbf{I}_2)$, we have that

$$P(\mathbf{T} \le \boldsymbol{\mu}_2(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})) = \Phi_2(\boldsymbol{\mu}_2(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})|\boldsymbol{\Gamma}\mathbf{b},\mathbf{I}_2) = \Phi_2(-\boldsymbol{\Gamma}\mathbf{b}|-\boldsymbol{\mu}_2(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}),\mathbf{I}_2),$$

which concludes the proof.

Proof of Theorem: From(A.1), (A.2) and (A.3), it follows that

$$\begin{aligned} f_{\mathbf{Y}}(\mathbf{y}|\boldsymbol{\theta},\boldsymbol{\lambda}) &= \int_{\mathbb{R}^q} 2^2\phi_n(\mathbf{y}|\mathbf{X}\boldsymbol{\beta},\boldsymbol{\Sigma})\phi_q(\mathbf{b}|\boldsymbol{\mu}_1(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}),\boldsymbol{\Lambda}) \\ &\quad \times \Phi_2(-\boldsymbol{\Gamma}\mathbf{b}|-\boldsymbol{\mu}_2(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}),\mathbf{I}_2)d\mathbf{b} \\ &= 2^2\phi_n(\mathbf{y}|\mathbf{X}\boldsymbol{\beta},\boldsymbol{\Sigma})E[\Phi_2(-\boldsymbol{\Gamma}\mathbf{W}|-\boldsymbol{\mu}_2(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}),\mathbf{I}_2)], \end{aligned}$$

where $\mathbf{W} \sim N_q(\boldsymbol{\mu}_1(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}),\boldsymbol{\Lambda})$. The proof is concluded by using Lemma **??**.

Proof of Corollary 2:

$$(\boldsymbol{\mu}_2-\boldsymbol{\Gamma}\boldsymbol{\mu}_1)(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) = \begin{pmatrix} \boldsymbol{\lambda}_e^T\boldsymbol{\psi}^{-1/2}(\boldsymbol{\psi}-\mathbf{Z}\boldsymbol{\Lambda}\mathbf{Z}^T)\boldsymbol{\psi}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) \\ \boldsymbol{\lambda}_b^T\mathbf{D}^{-1/2}\boldsymbol{\Lambda}\mathbf{Z}^T\boldsymbol{\psi}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) \end{pmatrix}$$

and

$$\mathbf{I}_2+\boldsymbol{\Gamma}\boldsymbol{\Lambda}\boldsymbol{\Gamma}^T = \begin{pmatrix} 1+\boldsymbol{\lambda}_e^T\boldsymbol{\psi}^{-1/2}\mathbf{Z}\boldsymbol{\Lambda}\mathbf{Z}^T\boldsymbol{\psi}^{-1/2}\boldsymbol{\lambda}_e & -\boldsymbol{\lambda}_e^T\boldsymbol{\psi}^{-1/2}\mathbf{Z}\boldsymbol{\Lambda}\mathbf{D}^{-1/2}\boldsymbol{\lambda}_b \\ -\boldsymbol{\lambda}_b^T\mathbf{D}^{-1/2}\boldsymbol{\Lambda}\mathbf{Z}^T\boldsymbol{\psi}^{-1/2}\boldsymbol{\lambda}_e & 1+\boldsymbol{\lambda}_b^T\mathbf{D}^{-1/2}\boldsymbol{\Lambda}\mathbf{D}^{-1/2}\boldsymbol{\lambda}_b \end{pmatrix},$$

which is a diagonal matrix when $\boldsymbol{\lambda}_e = \mathbf{0}$ or $\boldsymbol{\lambda}_b = \mathbf{0}$. Hence, for $\boldsymbol{\lambda}_e = \mathbf{0}$, the asymmetric part of (2.5) can be computed as

$$\begin{aligned} \Phi_2(\boldsymbol{\mu}_2-\boldsymbol{\Gamma}\boldsymbol{\mu}_1|\mathbf{0},\mathbf{I}_2+\boldsymbol{\Gamma}\boldsymbol{\Lambda}\boldsymbol{\Gamma}^T) &= \Phi_2((\mathbf{I}_2+\boldsymbol{\Gamma}\boldsymbol{\Lambda}\boldsymbol{\Gamma}^T)^{-1/2}(\boldsymbol{\mu}_2-\boldsymbol{\Gamma}\boldsymbol{\mu}_1)) \\ &= \frac{1}{2}\Phi_1\left(\frac{\boldsymbol{\lambda}_b^T\mathbf{D}^{-1/2}\boldsymbol{\Lambda}\mathbf{Z}^T\boldsymbol{\psi}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{\sqrt{1+\boldsymbol{\lambda}_b^T\mathbf{D}^{-1/2}\boldsymbol{\Lambda}\mathbf{D}^{-1/2}\boldsymbol{\lambda}_b}}\right), \end{aligned}$$

where some algebraic manipulations yield $\boldsymbol{\Lambda}\mathbf{Z}^T = \mathbf{D}\mathbf{Z}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\psi}$. Similarly, for $\boldsymbol{\lambda}_b = \mathbf{0}$, we have that

$$\Phi_2(\boldsymbol{\mu}_2-\boldsymbol{\Gamma}\boldsymbol{\mu}_1|\mathbf{0},\mathbf{I}_2+\boldsymbol{\Gamma}\boldsymbol{\Lambda}\boldsymbol{\Gamma}^T) = \frac{1}{2}\Phi_1\left(\frac{\boldsymbol{\lambda}_e^T\boldsymbol{\Psi}^{-1/2}(\boldsymbol{\psi}-\mathbf{Z}\boldsymbol{\Lambda}\mathbf{Z}^T)\boldsymbol{\psi}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{\sqrt{1+\boldsymbol{\lambda}_e^T\boldsymbol{\psi}^{-1/2}\mathbf{Z}\boldsymbol{\Lambda}\mathbf{Z}^T\boldsymbol{\psi}^{-1/2}\boldsymbol{\lambda}_e}}\right),$$

and the proof concludes by noting that $\boldsymbol{\psi}-\mathbf{Z}\boldsymbol{\Lambda}\mathbf{Z}^T = \boldsymbol{\psi}\boldsymbol{\Sigma}^{-1}\boldsymbol{\psi}$.

Proof of Proposition 2:

In fact, the joint density of $\mathbf{Y}$ and T is

$$f_{\mathbf{Y},T}(\mathbf{y},t|\boldsymbol{\theta},\boldsymbol{\lambda}) = 2\phi_n(\mathbf{y}|\boldsymbol{\mu}+\mathbf{d}t,\boldsymbol{\Psi})\phi_1(t)\mathbb{I}\{t>0\}.$$

After some simple algebraic manipulations using Lemma **??**, we have that

$$\phi_n(\mathbf{y}|\boldsymbol{\mu}+\mathbf{d}t,\boldsymbol{\Psi})\phi_1(t) = \phi_n(\mathbf{y}|\boldsymbol{\mu},\boldsymbol{\Sigma})\phi_1(t|\eta,\tau^2), \tag{A.6}$$

concluding the proof.

Lemma 4. Let $X \sim N(\eta, \tau^2)$. Then, for any real constant a it follows that

$$E[X|X > a] = \eta + \frac{\phi_1(\frac{a-\eta}{\tau})}{1 - \Phi_1(\frac{a-\eta}{\tau})}\tau,$$

$$E[X^2|X > a] = \eta^2 + \tau^2 + \frac{\phi_1(\frac{a-\eta}{\tau})}{1 - \Phi_1(\frac{a-\eta}{\tau})}(\eta + a)\tau.$$

Proof: See Johnson et al. (1994), Section 10.1.

Proof of Preposition 3:

Note that we can write

$$E(T^k|\mathbf{y}) = \int_{-\infty}^{\infty} t^k f(t|\mathbf{y})dt = \frac{1}{f_{\mathbf{Y}}(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\lambda})} \int_{-\infty}^{\infty} t^k f_{\mathbf{Y},T}(\mathbf{y}, t|\boldsymbol{\theta}, \boldsymbol{\lambda})dt.$$

From (4.1) and (4.3), it then follows that

$$E(T^k|\mathbf{y}) = \frac{1}{\Phi_1(\frac{\eta}{\tau})} \int_0^{\infty} t^k \phi_1(t|\eta, \tau^2)dt = E(X^k|X > 0),$$

where $X \sim N_1(\eta, \tau^2)$ and $\Phi_1(\frac{\eta}{\tau}) = P(X > 0)$. Thus, (4.4) and (4.5) follow from Lemma 4 with a=0, which concludes the proof.

References

Arellano-Valle, R. B., del Pino, G. and San Martin, E. (2002). Definition and probabilistic properties of skew distributions. *Statistic and Probability Letters* **58**, 111-121.

Arellano-Valle, R. B. and del Pino, G. (2003). From symmetric to asymmetric distributions: A unified approach. In *On Skew-distributions and Their Applications: A Journey Beyond Normality* (Edited by M. G. Genton). Chapman-Hall.

Arellano-Valle, R. B. and Genton, M. G. (2005). Fundamental skew distributions. *Journal of Multivariate Analysis* **96**, 93-116.

Arellano-Valle, R. B., Ozan, S., Bolfarine, H. and Lachos, V. H. (2003). Skew normal measurement error models. To appear, *Journal of Multivariate Analysis*.

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics* **12**, 171-178.

Azzalini, A. and Capitanio, A. (1999). Statistical applications of the multivariate skew normal distributions. *Journal Royal Statistics Society* **61**, 579-602.

Azzalini, A. and Dalla-Valle, A. (1996). The multivariate skew-normal distribution. *Biometrika* **83**, 715-726.

Box, G. E. P., Jenkins, G. M., and Reinsel, G. C. (1994). *Time Series Analysis: Forecasting and Control*(3rd ed.) Holden-Day.

Branco, M. and Dey, D. (2001). A general class of multivariate skew-elliptical distribution. *Journal of Multivariate Analysis* **79**, 93-113.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, ser. B* **39**, 1-22.

Gallant, A. R. and Nychka, D. W. (1987). Seminonparametric maximum likelihood estimation. *Econometrica* **55**, 363-390.

Genton, M. G. and Loperfido, N. (2001). Generalized skew-elliptical distributions and their quadratic forms. *Institute of Statistics Mimeo Series* #2541, under review.

Henze, N. (1986). A probabilistic representation of the "skew normal" distribution. *Scand. J. Statist.*, **13** 271-275.

Hu, P., Tsiatis, A. A. and Davidian, M. (1988). Estimating the parameters in the Cox models when covariate variables are measured with erro. *Biometrics* **54**, 1407-1419.

Johnson, N. L., Kotz, S. and Balakrishnan, N. (1994). *Continuous Univariate Distributions,* **Vol. 1.** Wiley.

Laird, N. M. and Ware, J. H. (1982). Random effects models for longitudinal data. *Biometrics* **38**, 963-974.

Lindstrom, M. J., and Bates, D. M. (1988). Newtom-Rhapson and EM Algorithm for linear mixed-effects model for repeated-measures. *Journal of the American Statistical Association* **83**, 1014-1022.

Mangder, L. S. and Zeger, S. L (1996). A smooth nonparametric estimate of a mixing distribution using mixtures of Gaussians. *Journal of the American Statistical Association* **91**, 1141-1151.

Meng, X. L. and Rubin, D.B.(1993). Maximum likelihood estimation via the ECM algorithm: A general Framework. *Biometrika* **80**, 267-278.

Pinheiro, J. C. and Bates, D.M.(2000). *Mixed-Effects Models in S and S-plus.* Springer-Verlag.

Sahu, S. K., Dey, D. K., and Branco, M. (2003). A new class of multivariate distributions with applications to Bayesian regression models. *The Canadian Journal of Statistics* **29**, 129-150.

Tao, H., Palta, M., Yandell, B. S. and Newton, M. A. (1999). An estimation method for the semiparametric mixed effects model. *Biometrics* **55**, 102-110.

Verbeke, G. and Lesaffre, E. (1996). A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association* **91**, 217-221.

Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. Springer.

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

Received June 29, 2004; accepted August 3, 2004.

Reinaldo B. Arellano-Valle
Departamento de Estadistica
Pontificia Universidad Catolica de Chile
Casilla 306, Correo 22, Santiago, Chile
reivalle@mat.puc.cl

Heleno Bolfarine
Instituto de Matematica e Estatistica
Universidade de Sao Paulo
Caixa Postal 6621-CEP 05315 970, Brasil
hbolfar@ime.usp.br

Victor H. Lachos
Instituto de Matematica e Estatistica
Universidade de Sao Paulo
Caixa Postal 6621-CEP 05315 970, Brasil
vhugo@ime.usp.br