Communications in Statistics - Theory and Methods
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/lsta20

Likelihood-Based Inference for Multivariate Skew-Normal Regression Models

Víctor H. Lachos [a] , Heleno Bolfarine [b] , Reinaldo B. Arellano-Valle [c] & Lourdes C. Montenegro [d]

[a] Departamento de Estatística , Universidade Estadual de Campinas , São Paulo, Brazil
[b] Departamento de Estatística , Universidade de São Paulo , São Paulo, Brazil
[c] Departamento de Estadística , Pontifícia Universidade Católica de Chile , Chile
[d] Departamento de Estatística , Universidade Federal de Minas Gerais , Minas Gerais, Brazil
Published online: 30 Jun 2007.

Communications in Statistics—Theory and Methods, 36: 1769–1786, 2007

ISSN: 0361-0926 print/1532-415X online
DOI: 10.1080/03610920601126241

Taylor & Francis
Taylor & Francis Group

Regression Analysis

Likelihood-Based Inference for Multivariate Skew-Normal Regression Models

VÍCTOR H. LACHOS[1], HELENO BOLFARINE[2], REINALDO B. ARELLANO-VALLE[3], AND LOURDES C. MONTENEGRO[4]

[1]Departamento de Estatística, Universidade Estadual de Campinas, São Paulo, Brazil
[2]Departamento de Estatística, Universidade de São Paulo, São Paulo, Brazil
[3]Departamento de Estadística, Pontifícia Universidade Católica de Chile, Chile
[4]Departamento de Estatística, Universidade Federal de Minas Gerais, Minas Gerais, Brazil

In this article, we present EM algorithms for performing maximum likelihood estimation for three multivariate skew-normal regression models of considerable practical interest. We also consider the restricted estimation of the parameters of certain important special cases of two models. The methodology developed is applied in the analysis of longitudinal data on dental plaque and cholesterol levels.

Keywords EM algorithm; Measurement error models; Mixed effects model; Skewness; Skew-normal distribution.

Mathematics Subject Classification Primary 62H12; Secondary 62E05.

1. Introduction

During the last decade, there has been a growing interest in the construction of flexible parametric classes of multivariate distributions exhibiting skewness. However, as pointed out in Azzalini (2005) (see Arnold and Beaver, 2002), the body of theoretical results that has been developed so far has not been widely applied to real problems. The practical usefulness of these stochastic models and their relevance in statistical applications is the main motivation for this work.

Received July 4, 2005; Accepted August 21, 2006
Address correspondence to Víctor H. Lachos, Departamento de Estatística, Universidade Estadual de Campinas, Caixa Postal 6065-CEP 13083-859, Campinas, SP, Brazil; E-mail: hlachos@ime.unicamp.br

A popular approach to constructing multivariate skewed normal distributions consists in modifying the probability density function (pdf) of a normal random vector in a multiplicative fashion. Although this idea has been present in the literature for some considerable time, it was Azzalini (1985, 1986) who thoroughly studied its use in the construction of univariate distributions, such as the so-called skew-normal distribution. The multivariate extension of the skew-normal distribution was subsequently considered by Azzalini and Dalla-Valle (1996), and its statistical application exhibited in Azzalini and Capitanio (1999). Unified generalizations have been developed recently by Arellano-Valle and Azzalini (2006) and Arellano-Valle and Genton (2005). A useful recent review of developments in this field is given by Azzalini (2005). See also the book edited by Genton (2004).

In this article, we consider a slightly modified version of the multivariate skew-normal distribution proposed by Azzalini and Dalla-Valle (1996), which is a special case of the fundamental skew-normal distribution proposed by Arellano-Valle and Genton (2005). As will be seen, this modification is advantageous in the sense that more efficient algorithms can be developed for obtaining the maximum likelihood (ML) estimates. Specifically, we say that an n-dimensional random vector $\mathbf{X}$ has a *standardized multivariate skew-normal distribution* with skewness parameter vector $\boldsymbol{\lambda}$, denoted by $\mathbf{X} \sim SN_n(\boldsymbol{\lambda})$, if its pdf is given by

$$f_{\mathbf{X}}(\mathbf{x}) = 2\phi_n(\mathbf{x})\Phi_1(\boldsymbol{\lambda}^\top \mathbf{x}), \quad \mathbf{x} \in \mathbb{R}^n, \tag{1}$$

where, as usual, $\phi_n(\cdot)$ and $\Phi_n(\cdot)$ denote, respectively, the pdf and cumulative distribution function (cdf) of the $N_n(\mathbf{0}, \mathbf{I}_n)$ distribution. More generally, for a $N_n(\boldsymbol{\mu}, \boldsymbol{\Psi})$ distribution, we denote the equivalent functions by $\phi_n(\cdot \mid \boldsymbol{\mu}, \boldsymbol{\Psi})$ and $\Phi_n(\cdot \mid \boldsymbol{\mu}, \boldsymbol{\Psi})$, respectively. As a location-scale extension of (1), we consider the distribution of $\mathbf{Y} = \boldsymbol{\mu} + \boldsymbol{\Psi}^{1/2}\mathbf{X}$, with pdf

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

We will use the notation $\mathbf{Y} \sim SN_n(\boldsymbol{\mu}, \boldsymbol{\Psi}, \boldsymbol{\lambda})$ to denote the multivariate skew-normal distribution followed by $\mathbf{Y}$. A stochastic representation is given by

$$\mathbf{Y} \stackrel{d}{=} \boldsymbol{\mu} + \boldsymbol{\Psi}^{1/2}(\boldsymbol{\delta}|T_0| + (\mathbf{I}_n - \boldsymbol{\delta}\boldsymbol{\delta}^\top)^{1/2}\mathbf{T}_1), \quad \text{with } \boldsymbol{\delta} = \frac{\boldsymbol{\lambda}}{\sqrt{1 + \boldsymbol{\lambda}^\top \boldsymbol{\lambda}}}, \tag{3}$$

where $|T_0|$ denotes the absolute value of T_0, $T_0 \sim N_1(0, 1)$, and $\mathbf{T}_1 \sim N_n(\mathbf{0}, \mathbf{I}_n)$ are independent, and "$\stackrel{d}{=}$" means "distributed as". For more details on this approach, see Arellano-Valle and Genton (2005) and Arellano-Valle et al. (2005).

Motivated also by the results for EM algorithm based estimation of the parameters of the skew-normal comparative calibration models developed in Lachos et al. (2005), in the following sections we present extensions to multivariate regression contexts, including measurement error models with null intercept and mixed models. We want to emphasize that the development of the models considered here arose out of the analysis of data sets, such as depicted in Fig. 1, for which there is pronounced evidence of skewness.

The remainder of the article is structured as follows. Section 2 deals with the case of independent observations from the multivariate skew-normal distribution, including the implementation of an EM-type algorithm for ML estimation which

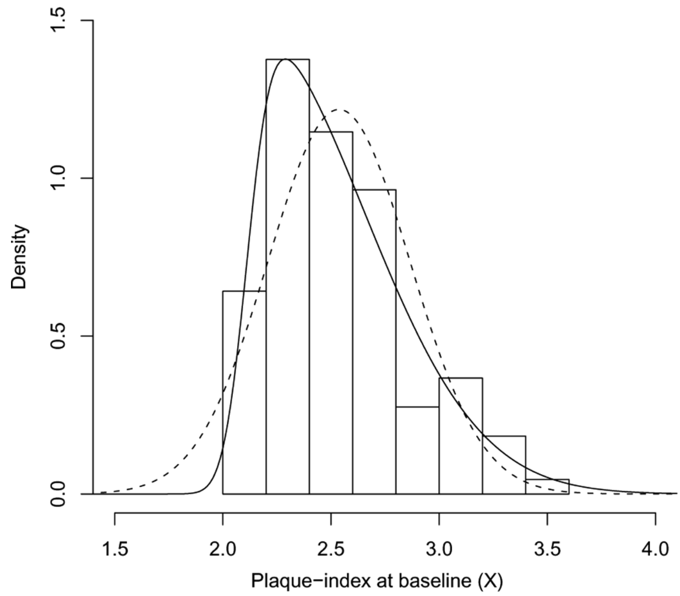

Figure 1. Dental plaque index data set. Histogram of the observed covariate x (plaque index measured at the start of the study) with superimposed fitted skew-normal (solid) and normal (dashed) densities.

yields closed form expressions for the equations in the M-step. In Sec. 3, we study a multivariate skew-normal null intercept measurement error model (SN-IMEM) with a dependence structure between the response variables within the same group appropriate for longitudinal data analysis. The EM algorithm is used to perform estimation for both restricted and unrestricted models. In Sec. 4, the skew-normal linear mixed model (SN-LMM) is defined and we describe an approach for restricted parameter estimation based on the EM algorithm. The methodology proposed for the SN-IMEM and SN-LMM models is illustrated in Sec. 5 using two real data sets. Some concluding remarks are presented in Sec. 6.

2. Multivariate Skew-Normal Responses

Suppose that we have observations on m independent individuals, $\mathbf{y}_1, \ldots, \mathbf{y}_m$, where $\mathbf{y}_i \sim SN_n(\boldsymbol{\mu}_i, \boldsymbol{\Psi}, \boldsymbol{\lambda})$, $i = 1, \ldots, m$. Associated with individual i we assume a known $n \times p$ covariate matrix $\mathbf{X}_i$, which we use to specify the linear predictor $\boldsymbol{\mu}_i = \mathbf{X}_i\boldsymbol{\beta}$, where $\boldsymbol{\beta}$ is a p-dimensional vector of unknown regression coefficients. Under these conditions, the log-likelihood function can be written as

$$\ell(\boldsymbol{\theta}) = \log 2 - \frac{1}{2} m \log|\boldsymbol{\Psi}| - \frac{1}{2}\sum_{i=1}^{m}(\mathbf{y}_i - \boldsymbol{\mu}_i)^\top \boldsymbol{\Psi}^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_i) + \sum_{i=1}^{m} \log \Phi_1\big(\boldsymbol{\lambda}^\top \boldsymbol{\Psi}^{-1/2}(\mathbf{y}_i - \boldsymbol{\mu}_i)\big).$$

As in Azzalini and Capitanio (1999), it is possible to find the ML estimate of the parameter vector $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \boldsymbol{\alpha}^\top, \boldsymbol{\lambda}^\top)^\top$, where $\boldsymbol{\alpha}$ denotes a minimal set of parameters such that $\boldsymbol{\Psi}$ is well defined (e.g., the upper triangular elements of $\boldsymbol{\Psi}$ in the unstructured case), by directly maximizing the log-likelihood. This can be done, for instance, using Azzalini's library from his home page. However, we prefer using the EM algorithm which, as mentioned in his library, although quite slow seems to be very robust.

In what follows, we use the EM algorithm in conjunction with the marginal stochastic representation given in (3) to obtain the ML estimate of $\boldsymbol{\theta}$. We note that it is hard to implement this approach for the multivariate skew-normal distribution proposed by Azzalini and Dalla-Valle (1996) due to its conditional stochastic representation.

2.1. *The EM Algorithm*

The EM algorithm is a popular iterative algorithm for ML estimation for models with incomplete data. More specifically, let $\mathbf{y}$ denote the observed data and $\mathbf{s}$ denote the missing data. The complete data $\mathbf{y}_c = (\mathbf{y}, \mathbf{s})$ is $\mathbf{y}$ augmented with $\mathbf{s}$. We denote by $\ell_c(\boldsymbol{\theta} \mid \mathbf{y}_c)$, $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, the complete-data log-likelihood function and by $Q(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}}) = E[\ell_c(\boldsymbol{\theta} \mid \mathbf{y}_c) \mid \mathbf{y}, \hat{\boldsymbol{\theta}}]$, the expected complete-data log-likelihood function. Each iteration of the EM algorithm involves two steps: an E-step and an M-step, defined as:

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

Notice that, by using (3), the set-up defined above can be written as

$$\mathbf{Y}_i \mid T_i = t_i \overset{\text{ind}}{\sim} N_n(\mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\Psi}^{1/2}\boldsymbol{\delta}t_i, \boldsymbol{\Psi}^{1/2}(\mathbf{I}_k - \boldsymbol{\delta}\boldsymbol{\delta}^\top)\boldsymbol{\Psi}^{1/2}), \tag{4}$$

$$T_i \overset{\text{iid}}{\sim} HN_1(0, 1) \quad i = 1, \ldots, m, \tag{5}$$

all independent, where $HN_1(0, 1)$ denotes the univariate standard half-normal distribution (see $|T_0|$ in Eq. (3) or Johnson et al., 1994).

Let $\mathbf{y} = (\mathbf{y}_1^\top, \ldots, \mathbf{y}_m^\top)^\top$ and $\mathbf{t} = (t_1, \ldots, t_m)^\top$. Then, under the hierarchical representation (4)–(5), with $\boldsymbol{\Delta} = \boldsymbol{\Psi}^{1/2}\boldsymbol{\delta}$ and $\boldsymbol{\Gamma} = \boldsymbol{\Psi} - \boldsymbol{\Delta}\boldsymbol{\Delta}^\top$, it follows that the complete log-likelihood function associated with $\mathbf{y}_c = (\mathbf{y}^\top, \mathbf{t}^\top)^\top$ is

$$\ell_c(\boldsymbol{\theta} \mid \mathbf{y}_c) = -\frac{m}{2}\log|\boldsymbol{\Gamma}| - \frac{1}{2}\sum_{i=1}^{m}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \boldsymbol{\Delta}t_i)^\top\boldsymbol{\Gamma}^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \boldsymbol{\Delta}t_i) - \frac{1}{2}\sum_{i=1}^{m} t_i^2 + C, \tag{6}$$

where C is a constant that is independent of the parameter vector $\boldsymbol{\theta}$. Letting $\hat{t}_i = \mathrm{E}[T_i \mid \boldsymbol{\theta} = \hat{\boldsymbol{\theta}}, \mathbf{y}_i]$ and $\hat{t}_i^2 = \mathrm{E}[T_i^2 \mid \boldsymbol{\theta} = \hat{\boldsymbol{\theta}}, \mathbf{y}_i]$, we obtain, using the moments of the truncated normal distribution (see Arellano-Valle et al., 2005), that

$$\hat{t}_i = \hat{\mu}_{Ti} + W_{\Phi_1}\left(\frac{\hat{\mu}_{T_i}}{\widehat{M}_{T_i}}\right)\widehat{M}_{T_i} \quad \text{and} \quad \hat{t}_i^2 = \hat{\mu}_{T_i}^2 + \widehat{M}_{T_i}^2 + W_{\Phi_1}\left(\frac{\hat{\mu}_{T_i}}{\widehat{M}_{T_i}}\right)\widehat{M}_{T_i}\hat{\mu}_{T_i}, \tag{7}$$

where $W_{\Phi_1}(u) = \phi_1(u)/\Phi_1(u)$, $\widehat{M}_{T_i}^2 = 1/(1 + \widehat{\boldsymbol{\Delta}}^\top\widehat{\boldsymbol{\Gamma}}^{-1}\widehat{\boldsymbol{\Delta}})$, and $\hat{\mu}_{T_i} = \widehat{\boldsymbol{\Delta}}^\top\widehat{\boldsymbol{\Gamma}}^{-1}(\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}})/(1 + \widehat{\boldsymbol{\Delta}}^\top\widehat{\boldsymbol{\Gamma}}^{-1}\widehat{\boldsymbol{\Delta}})$, $i = 1, \ldots, m$.

It follows, after some simple algebra and using (7), that the expectation with respect to $\mathbf{t}$ conditional on $\mathbf{y}$, of the complete log-likelihood function, has the form

$$Q(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}}) = \mathrm{E}[\ell_c(\boldsymbol{\theta} \mid \mathbf{y}_c) \mid \mathbf{y}, \hat{\boldsymbol{\theta}}] = -\frac{m}{2}\log|\boldsymbol{\Gamma}| - \frac{1}{2}\sum_{i=1}^{m}\big(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \boldsymbol{\Delta}\hat{t}_i\big)^\top\boldsymbol{\Gamma}^{-1}\big(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \boldsymbol{\Delta}\hat{t}_i\big)$$
$$- \frac{1}{2}\sum_{i=1}^{m}(\hat{t}_i^2 - (\hat{t}_i)^2)\boldsymbol{\Delta}^\top\boldsymbol{\Gamma}^{-1}\boldsymbol{\Delta} - \frac{1}{2}\sum_{i=1}^{m}\hat{t}_i^2 + C. \tag{8}$$

We then have the following EM algorithm:

E-step: Given $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$, compute $\hat{t}_i$, $\hat{t}_i^2$ for $i = 1, \ldots, m$ using (7).

M-step: Update $\hat{\boldsymbol{\theta}}$ by maximizing $Q(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}})$ over $\boldsymbol{\theta}$, which leads to the following closed form expressions

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{m}\mathbf{X}_i^\top\boldsymbol{\Gamma}^{-1}\mathbf{X}_i\right)^{-1}\sum_{i=1}^{m}\mathbf{X}_i\boldsymbol{\Gamma}^{-1}\big(\mathbf{y}_i - \boldsymbol{\Delta}\hat{t}_i\big),$$
$$\widehat{\boldsymbol{\Gamma}} = \frac{1}{m}\sum_{i=1}^{m}\big[\big(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \boldsymbol{\Delta}\hat{t}_i\big)\big(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \boldsymbol{\Delta}\hat{t}_i\big)^\top + \big(\hat{t}_i^2 - (\hat{t}_i)^2\big)\boldsymbol{\Delta}\boldsymbol{\Delta}^\top\big], \tag{9}$$
$$\widehat{\boldsymbol{\Delta}} = \frac{\sum_{i=1}^{m}\hat{t}_i(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})}{\sum_{i=1}^{m}\hat{t}_i^2}.$$

The skewness parameter vector and the unstructured scale matrix can be estimated by noting that $\widehat{\boldsymbol{\Psi}} = \widehat{\boldsymbol{\Gamma}} + \widehat{\boldsymbol{\Delta}}\widehat{\boldsymbol{\Delta}}^T$ and $\hat{\boldsymbol{\lambda}} = \widehat{\boldsymbol{\Psi}}^{-1/2}\widehat{\boldsymbol{\Delta}}/(1 - \widehat{\boldsymbol{\Delta}}^\top\widehat{\boldsymbol{\Psi}}^{-1}\widehat{\boldsymbol{\Delta}})^{1/2}$. It is clear that when $\boldsymbol{\lambda} = \mathbf{0}$ (or $\boldsymbol{\Delta} = \mathbf{0}$) the M-step equations reduce to the equations obtained assuming normality. Starting values required to implement this algorithm and some inferential strategies will be discussed in the next sections.

3. Multivariate Skew-Normal Null Intercept Measurement Error Models

Measurement error models constitute an attractive option for modeling many practical experimental problems, especially when the same response is observed for the same individuals under different experimental conditions. One special case is the situation in which the same experimental unit is measured using two or more measuring devices. An extensive bibliography for such models can be found in Fuller (1987) and Cheng and Van Ness (1999). Recently, Aoki et al. (2003) discussed a multivariate normal (symmetric) null intercept measurement error model (N-IMEM), with a dependence structure between the response variables within the same group, applicable to longitudinal data studies. Although the normality assumption (or symmetry) is reasonable in many situations, it is not appropriate when the data exhibit non-normal behavior such as asymmetry. This is the case, for instance, with the dental clinical trial data set (Hadgu and Koch, 1999) where the observed dental plaque index at the start of the clinical trial may require transformation in order to be better approximated by the normal distribution. In this section, we extend the previously proposed multivariate normal model by considering that the true dental plaque index at the start of the trial (a latent variable) follows a skew-normal distribution, that is,

$$\begin{aligned} \mathbf{x}_i &= \boldsymbol{\xi}_i + \boldsymbol{\delta}_i, \\ \mathbf{y}_i &= \mathbf{X}_i\boldsymbol{\beta}_i + \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, m, \end{aligned} \tag{10}$$

where $\mathbf{x}_i = (x_{i1}, \ldots, x_{in_i})^\top$, $\mathbf{y}_i = (\mathbf{y}_{1i}^\top, \mathbf{y}_{2i}^\top)^\top = (y_{1i1}, \ldots, y_{1in_i}, y_{2i1}, \ldots, y_{2in_i})^\top$, $\boldsymbol{\beta}_i = (\beta_{1i}, \beta_{2i})^\top$, $\boldsymbol{\xi}_i = (\xi_{i1}, \ldots, \xi_{in_i})^\top$, $\boldsymbol{\epsilon} = (\boldsymbol{\epsilon}_{1i}^\top, \boldsymbol{\epsilon}_{2i}^\top)^\top = (\epsilon_{1i1}, \ldots, \epsilon_{1in_i}, \epsilon_{2i1}, \ldots, \epsilon_{2in_i})^\top$, $\boldsymbol{\delta}_i = (\delta_{i1}, \ldots, \delta_{in_i})^\top$, with $\delta_{ij} \overset{\text{iid}}{\sim} N_1(0, \sigma^2)$, $\epsilon_{kij} \overset{\text{ind}}{\sim} N_1(0, \sigma^2_{\epsilon i})$, δ_{ij} and ϵ_{kij} uncorrelated and independent of $\xi_{ij} \overset{\text{iid}}{\sim} SN_1(\mu_x, \sigma_x^2, \lambda_x)$, $i = 1, \ldots, m$, $j = 1, \ldots, n_i$, $k = 1, 2$, and

$$\mathbf{X}_i = \begin{bmatrix} \boldsymbol{\xi}_i & 0 \\ 0 & \boldsymbol{\xi}_i \end{bmatrix}.$$

We refer to this model as the skew-normal null intercept measurement error model (SN-IMEM). It contains, as a special case, the normal null intercept measurement error model (N-IMEM). Notice that the $\mathbf{X}_i$'s are unobservable matrices.

Now, define by $\mathbf{z}_{ij} = (y_{1ij}, y_{2ij}, x_{ij})^\top$ the observed vector. Then, using the stochastic representation in (3) with $n = 1$, we can write the model defined in (10) as

$$\mathbf{z}_{ij} \mid \xi_{ij} \overset{\text{ind}}{\sim} N_3(\xi_{ij}\boldsymbol{\beta}_{oi}, D(\boldsymbol{\phi}_i)), \tag{11}$$

$$\xi_{ij} \mid T_{ij} = t_{ij} \overset{\text{ind}}{\sim} N_1(\mu_x + \sigma_x\delta_x t_{ij}, \quad \sigma_x^2(1 - \delta_x^2)), \tag{12}$$

$$T_{ij} \overset{\text{iid}}{\sim} HN_1(0, 1), \tag{13}$$

$i = 1, \ldots, m$ and $j = 1, \ldots, n_i$, all independent, where $\boldsymbol{\beta}_{oi} = (\boldsymbol{\beta}_i^\top, 1)^\top = (\beta_{1i}, \beta_{2i}, 1)^\top$, $D(\boldsymbol{\phi}_i) = \text{diag}(\boldsymbol{\phi}_i)$, $\boldsymbol{\phi}_i = (\sigma^2_{\epsilon i}, \sigma^2_{\epsilon i}, \sigma^2)^\top$, and $\delta_x = \lambda_x/(1 + \lambda_x^2)^{1/2}$. For this type of model, classical inference for the parameter vector $\boldsymbol{\theta} = (\mu_x, \sigma_x^2, \lambda_x, \boldsymbol{\beta}^\top, \sigma^2, \boldsymbol{\sigma}^2_\epsilon)^\top$, with $\boldsymbol{\beta} = (\beta_{11}, \ldots, \beta_{1m}, \beta_{21}, \ldots, \beta_{2m})^\top$ and $\boldsymbol{\sigma}^2_\epsilon = (\sigma^2_{\epsilon 1}, \ldots, \sigma^2_{\epsilon m})^\top$, is based on the marginal distribution for the response $\mathbf{z}_{ij}$, which, after some algebra, can be written as

$$f(\mathbf{z}_{ij} \mid \boldsymbol{\theta}) = 2\phi_3(\mathbf{z}_{ij} \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)\Phi_1\big(\bar{\boldsymbol{\lambda}}_{xi}^\top \boldsymbol{\Sigma}_i^{-1/2}(\mathbf{z}_{ij} - \boldsymbol{\mu}_i)\big), \tag{14}$$

$i = 1, \ldots, m$ and $j = 1, \ldots, n_i$, i.e., $\mathbf{z}_{ij} \overset{\text{ind}}{\sim} SN_3(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \bar{\boldsymbol{\lambda}}_{xi})$, with $\boldsymbol{\mu}_i = \boldsymbol{\beta}_{oi}\mu_x$, $\boldsymbol{\Sigma}_i = D(\boldsymbol{\phi}_i) + \sigma_x^2\boldsymbol{\beta}_{oi}\boldsymbol{\beta}_{oi}^\top$, $\bar{\boldsymbol{\lambda}}_{xi} = \lambda_x\sigma_x^2\boldsymbol{\Sigma}_i^{-1/2}\boldsymbol{\beta}_{oi}/\sqrt{\sigma_x^2 + \lambda_x^2\boldsymbol{\Lambda}_i}$, where $\boldsymbol{\Lambda}_i = (\sigma_x^{-2} + \boldsymbol{\beta}_{oi}^\top D^{-1}(\boldsymbol{\phi}_i)\boldsymbol{\beta}_{oi})^{-1}$. It follows that the log-likelihood function for $\boldsymbol{\theta}$, given the observed sample $\mathbf{z} = (\mathbf{z}_1^\top, \ldots, \mathbf{z}_m^\top)^\top$, with $\mathbf{z}_i = (\mathbf{z}_{i1}^\top, \ldots, \mathbf{z}_{in_i}^\top)^\top$ can be written as $\ell(\boldsymbol{\theta}) = \sum_{i=1}^m \sum_{j=1}^{n_i} \ell_{ij}(\boldsymbol{\theta})$, where

$$\ell_{ij}(\boldsymbol{\theta}) = \log(2) - \frac{n_i}{2}\log(2\pi) - \frac{1}{2}(\mathbf{z}_{ij} - \boldsymbol{\mu}_i)^\top\boldsymbol{\Sigma}_i^{-1}(\mathbf{z}_{ij} - \boldsymbol{\mu}_i) + \log(K_{ij}),$$

with $K_{ij} = \Phi_1(\bar{\boldsymbol{\lambda}}_{xi}^\top\boldsymbol{\Sigma}_i^{-1/2}(\mathbf{z}_{ij} - \boldsymbol{\mu}_i))$.

As this log-likelihood function can be easily evaluated with high precision, model selection procedures based on criteria such as the Akaike Information Criterion (AIC), Schwarz's Bayesian Information Criterion (BIC), and the Hannan-Quinn Criterion (HQ) can be implemented. The asymptotic covariance matrix of the ML estimator can be estimated using the Hessian matrix which can be computed numerically using, for instance, the *optim* routine in *R*, *Jacobian* in Matlab, or *Num2derivate* in Ox. The observed information matrix can also be derived algebraically (see, for example, Lachos et al., 2005). In the next section we describe an EM algorithm which can be used to obtain the ML estimate of the parameter vector $\boldsymbol{\theta}$.

3.1. *An EM Algorithm for ML Estimation*

Let $\mathbf{z}_c = (\mathbf{z}^\top, \boldsymbol{\xi}^\top, \mathbf{t}^\top)^\top$, with $\mathbf{z} = (\mathbf{z}_1^\top, \ldots, \mathbf{z}_m^\top)^\top$, $\boldsymbol{\xi} = (\boldsymbol{\xi}_1^\top, \ldots, \boldsymbol{\xi}_m^\top)^\top$, and $\mathbf{t} = (\mathbf{t}_1^\top, \ldots, \mathbf{t}_m^\top)^\top$, where $\mathbf{z}_i = (\mathbf{z}_{i1}^\top, \ldots, \mathbf{z}_{in_i}^\top)^\top$, $\boldsymbol{\xi}_i = (\xi_{i1}, \ldots, \xi_{in_i})^\top$, and $\mathbf{t}_i = (t_{i1}, \ldots, t_{in_i})^\top$, $i = 1, \ldots, m$. In the following we implement the EM algorithm using double augmentation, by considering that $(\boldsymbol{\xi}, \mathbf{t})$ are missing data. Hence, under the hierarchical representation (11)–(13), with $\nu^2 = \sigma_x^2(1 - \delta_x^2)$ and $\varsigma = \sigma_x \delta_x$, it follows that the complete log-likelihood function associated with $\mathbf{z}_c$ is

$$\ell_c(\boldsymbol{\theta} \mid \mathbf{z}_c) = -\sum_{i=1}^{m} n_i \log \sigma_{\epsilon i}^2 - \frac{N}{2}\log\sigma^2 - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n_i}(\mathbf{z}_{ij} - \xi_{ij}\boldsymbol{\beta}_{oi})^\top D^{-1}(\boldsymbol{\phi}_i)(\mathbf{z}_{ij} - \xi_{ij}\boldsymbol{\beta}_{oi})$$
$$- \frac{N}{2}\log \nu^2 - \frac{1}{2\nu^2}\sum_{i=1}^{m}\sum_{j=1}^{n_i}(\xi_{ij} - \mu_x - \varsigma t_{ij})^2 - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n_i} t_{ij}^2 + C,$$

where $N = \sum_{i=1}^m n_i$ and C is a constant that is independent of the parameter vector $\boldsymbol{\theta}$. Letting $\hat{\xi}_{ij} = \mathrm{E}[\xi_{ij} \mid \boldsymbol{\theta} = \hat{\boldsymbol{\theta}}, \mathbf{z}_{ij}]$, $\hat{\xi^2}_{ij} = \mathrm{E}[\xi_{ij}^2 \mid \boldsymbol{\theta} = \hat{\boldsymbol{\theta}}, \mathbf{z}_{ij}]$, $\hat{t}_{ij} = \mathrm{E}[T_{ij} \mid \boldsymbol{\theta} = \hat{\boldsymbol{\theta}}, \mathbf{z}_{ij}]$, $\hat{t^2}_{ij} = \mathrm{E}[T_{ij}^2 \mid \boldsymbol{\theta} = \hat{\boldsymbol{\theta}}, \mathbf{z}_{ij}]$, and $\widehat{t\xi}_{ij} = \mathrm{E}[T_{ij}\xi_{ij} \mid \boldsymbol{\theta} = \hat{\boldsymbol{\theta}}, \mathbf{z}_{ij}]$, we obtain, using double conditional expectations and the moments of the truncated normal distribution, that

$$\hat{t}_{ij} = \hat{\mu}_{T_{ij}} + W_{\Phi_1}\left(\frac{\hat{\mu}_{T_{ij}}}{\widehat{M}_{T_i}}\right)\widehat{M}_{T_i}, \quad \hat{t^2}_{ij} = \hat{\mu}_{T_{ij}}^2 + \widehat{M}_{T_i}^2 + W_{\Phi_1}\left(\frac{\hat{\mu}_{T_{ij}}}{\widehat{M}_{T_i}}\right)\widehat{M}_{T_i}\hat{\mu}_{T_{ij}},$$
$$\hat{\xi}_{ij} = \hat{r}_{ij} + \hat{s}_i\, \hat{t}_{ij}, \quad \hat{\xi^2_{ij}} = \widehat{T}_{x_i}^2 + \hat{r}_{ij}^2 + 2\hat{r}_{ij}\, \hat{s}_i\, \hat{t}_{ij} + \hat{s}_i^{\,2}\, \hat{t^2_{ij}}, \quad \widehat{t\xi}_{ij} = \hat{r}_{ij}\hat{t}_{ij} + \hat{s}_i\, \hat{t^2_{ij}}, \quad (15)$$

where $W_{\Phi_1}(u) = \phi_1(u)/\Phi_1(u)$, $\widehat{M}_{T_i}^2 = [1 + \hat{\varsigma}^2\hat{\boldsymbol{\beta}}_{oi}^\top(D(\hat{\boldsymbol{\phi}}_i) + \hat{\nu}^2\hat{\boldsymbol{\beta}}_{oi}\hat{\boldsymbol{\beta}}_{oi}^\top)^{-1}\hat{\boldsymbol{\beta}}_{oi}]^{-1}$, $\hat{\mu}_{T_{ij}} = \hat{\varsigma}\widehat{M}_{T_i}^2\hat{\boldsymbol{\beta}}_{oi}^\top(D(\hat{\boldsymbol{\phi}}_i) + \hat{\nu}^2\hat{\boldsymbol{\beta}}_{oi}\hat{\boldsymbol{\beta}}_{oi}^\top)^{-1}(\mathbf{z}_{ij} - \hat{\boldsymbol{\beta}}_{oi}\hat{\mu}_x)$, $\widehat{T}_{x_i}^2 = \hat{\nu}^2[1 + \hat{\nu}^2\hat{\boldsymbol{\beta}}_{oi}^\top D^{-1}(\hat{\boldsymbol{\phi}}_i)\hat{\boldsymbol{\beta}}_{oi}]^{-1}$, $\hat{r}_{ij} = \hat{\mu}_x + \widehat{T}_{x_i}^2\hat{\boldsymbol{\beta}}_{oi}^\top D^{-1}(\hat{\boldsymbol{\phi}}_i)(\mathbf{z}_{ij} - \hat{\boldsymbol{\beta}}_{oi}\hat{\mu}_x)$, and $\hat{s}_i = \hat{\varsigma}(1 - \widehat{T}_{x_i}^2\hat{\boldsymbol{\beta}}_{oi}^\top D^{-1}(\hat{\boldsymbol{\phi}}_i)\hat{\boldsymbol{\beta}}_{oi})$.

Computing the expectation of $\ell_c(\boldsymbol{\theta} \mid \mathbf{z}_c)$ with respect to $(\boldsymbol{\xi}, \mathbf{t})$ conditional on $\mathbf{z}$, we obtain

$$Q(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}}) = E[\ell_c(\boldsymbol{\theta} \mid \mathbf{z}_c) \mid \mathbf{z}, \hat{\boldsymbol{\theta}}]$$
$$= -\sum_{i=1}^{m} n_i \log\sigma_{\epsilon i}^2 - \frac{N}{2}\log\sigma^2 - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n_i}(\mathbf{z}_{ij} - \boldsymbol{\beta}_{oi}\hat{\xi}_{ij})^\top D^{-1}(\boldsymbol{\phi}_i)(\mathbf{z}_{ij} - \boldsymbol{\beta}_{oi}\hat{\xi}_{ij})$$
$$- \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n_i}(\hat{\xi^2_{ij}} - (\hat{\xi}_{ij})^2)\boldsymbol{\beta}_{oi}^\top \mathrm{D}^{-1}(\phi)\boldsymbol{\beta}_{oi}$$

$$-\frac{N}{2}\log(\nu^2)-\frac{1}{2\nu^2}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(\hat{\xi}^2_{ij}+\mu_x^2+\varsigma^2\hat{t}^2_{ij}-2\hat{\xi}_{ij}\mu_x-2\varsigma\widehat{t\xi}_{ij}+2\varsigma\mu_x\hat{t}_{ij}\right)$$

$$-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\hat{t}^2_{ij}+C.$$

Therefore, the following EM algorithm can be used to obtain the ML estimate of $\boldsymbol{\theta}$ for the model defined in (10):

E-step: Given $\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}$, compute $\hat{t}_{ij}$, $\hat{t}^2_{ij}$, $\hat{\xi}_{ij}$, $\hat{\xi}^2_{ij}$, and $\widehat{t\xi}_{ij}$ for $i=1,\ldots,m$ and $j=1,\ldots,n_i$ using (15).

M-step: Update $\hat{\boldsymbol{\theta}}$ by maximizing $Q(\boldsymbol{\theta}\,|\,\hat{\boldsymbol{\theta}})$ over $\boldsymbol{\theta}$, which leads to

$$\hat{\beta}_{ki}=\frac{\sum_{j=1}^{n_i}y_{kij}\hat{\xi}_{ij}}{\sum_{j=1}^{n_i}\hat{\xi}^2_{ij}},\quad k=1,2,\ i=1,\ldots,m, \tag{16}$$

$$\hat{\mu}_x=\frac{1}{N}\sum_{i=1}^{m}\sum_{j=1}^{n_i}(\hat{\xi}_{ij}-\varsigma\hat{t}_{ij}), \tag{17}$$

$$\hat{\sigma}^2_{\epsilon i}=\frac{1}{2n_i}\sum_{j=1}^{n_i}\left(y^2_{1ij}+y^2_{2ij}-2\hat{\xi}_{ij}(y_{1ij}\beta_{1i}+y_{2ij}\beta_{2i})+\hat{\xi}^2_{ij}\left(\beta^2_{1i}+\beta^2_{2i}\right)\right),\quad i=1,\ldots,m, \tag{18}$$

$$\hat{\sigma}^2=\frac{1}{N}\sum_{i=1}^{m}\sum_{j=1}^{n_i}(x^2_{ij}-2\hat{\xi}_{ij}x_{ij}+\hat{\xi^2}_{ij}),\quad \hat{\varsigma}=\frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i}(\widehat{t\xi}_{ij}-\mu_x\hat{t}_{ij})}{\sum_{i=1}^{m}\sum_{j=1}^{n_i}\hat{t}^2_{ij}}, \tag{19}$$

$$\hat{\nu}^2=\frac{1}{N}\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(\hat{\xi}^2_{ij}+\mu_x^2+\varsigma^2\hat{t}^2_{ij}-2\mu_x\hat{\xi}_{ij}-2\varsigma\widehat{t\xi}_{ij}+2\varsigma\mu_x\hat{t}_{ij}\right). \tag{20}$$

Note that (16) can be written, alternatively, as $\hat{\boldsymbol{\beta}}=\boldsymbol{\Delta}_{\xi}^{-1}\boldsymbol{\eta}_{\xi y}$, where $\boldsymbol{\Delta}_{\xi}=\mathrm{diag}(\sum_{j=1}^{n_1}\hat{\xi}^2_{1j},\ldots,\sum_{j=1}^{n_m}\hat{\xi}^2_{mj},\sum_{j=1}^{n_1}\hat{\xi}^2_{1j},\ldots,\sum_{j=1}^{n_m}\hat{\xi}^2_{mj})$ and $\boldsymbol{\eta}_{\xi y}=(\sum_{j=1}^{n_1}y_{11j}\hat{\xi}_{1j},\ldots,\sum_{j=1}^{n_m}y_{1mj}\hat{\xi}_{mj},\sum_{j=1}^{n_1}y_{21j}\hat{\xi}_{1j},\sum_{j=1}^{n_1}\ldots,\sum_{j=1}^{n_m}y_{2mj}\hat{\xi}_{mj})^{\top}$. The shape and scale parameters of the latent variable $\boldsymbol{\xi}$ can be estimated by noting that $\varsigma/\nu=\lambda_x$, and $\phi_x=\varsigma^2+\nu^2$. Useful starting values are those obtained under the normality assumption, with the starting values for the skewness parameters set equal to 0. However, in order to ensure that the true ML estimate is identified, we recommend running the EM algorithm using a range of different starting values.

3.2. *Restricted Estimation*

Two of the main objectives in analyses based on this type of model are to estimate $\boldsymbol{\beta}$ and to test hypotheses such as $H_{01}:\beta_{k1}=\cdots=\beta_{km}$ and $H_{02}:\beta_{1i}=\beta_{2i}$, $k=1,2$, $i=1,\ldots,m$. Thus, it is also important to identify estimation procedures under the restriction imposed by the null hypothesis, and to use these results to construct, for example, likelihood ratio statistics ($\daleth_{LR}$) for testing the hypotheses. Suppose, then, that our interest centers on estimating the parameter $\boldsymbol{\beta}$ under q linearly independent restrictions defined as $\mathbf{C}_s^{\top}\boldsymbol{\beta}-d_s=0$, where $\mathbf{C}_s$, $s=1,\ldots,q$, are $2m\times 1$ vectors and d_s, $s=1,\ldots,q$, are scalars, both of which are assumed to be known and fixed. In the M-step of the EM algorithm, the problem is then to maximize the conditional expectation of the complete log-likelihood function $E[\ell_c(\boldsymbol{\theta}\mid\mathbf{z}_c)\mid\mathbf{z},\hat{\boldsymbol{\theta}}]$ subject to the linear constraints $\mathbf{C}\boldsymbol{\beta}-\mathbf{d}=\mathbf{0}$, where $\mathbf{C}=(\mathbf{C}_1,\ldots,\mathbf{C}_q)^\top$ and $\mathbf{d}=(d_1,\ldots,d_q)^\top$. Similarly, as in Nyquist (1991), we apply the methodology of penalty functions and consider the quadratic penalty function

$$P(\boldsymbol{\theta},\boldsymbol{\Upsilon})=E[\ell_c(\boldsymbol{\theta}\mid\mathbf{z}_c)\mid\mathbf{z},\hat{\boldsymbol{\theta}}]-\frac{1}{2}\sum_{s=1}^{q}\gamma_s(d_s-\mathbf{C}_s^\top\boldsymbol{\beta})^2,$$

with $\boldsymbol{\Upsilon}=(\gamma_1,\ldots,\gamma_q)$. The procedure consists in finding the solution of $\max_{\boldsymbol{\theta}}P(\boldsymbol{\theta},\boldsymbol{\Upsilon})$ for positive and fixed values of γ_s, $s=1,\ldots,q$. The solution for $\boldsymbol{\beta}$ will be denoted by $\boldsymbol{\beta}(\boldsymbol{\Upsilon})$. The equality restricted estimate $\boldsymbol{\beta}$ is given by

$$\tilde{\boldsymbol{\beta}}_c=\lim_{\gamma_1,\ldots,\gamma_q\to\infty}\boldsymbol{\beta}(\boldsymbol{\Upsilon}).$$

Using a similar approach to that given in Nyquist (1991), one may show that $\boldsymbol{\beta}(\boldsymbol{\Upsilon})$ is the solution of the following iterative process:

$$\begin{aligned}\tilde{\boldsymbol{\beta}}_c^{(r+1)}&=\boldsymbol{\Delta}_\xi^{(r)^{-1}}\boldsymbol{\eta}_{\xi y}^{(r)}+\boldsymbol{\Delta}_\xi^{(r)^{-1}}\mathbf{C}^\top\big[\mathbf{C}\boldsymbol{\Delta}_\xi^{(r)^{-1}}\mathbf{C}^\top\big]^{-1}\big[\mathbf{d}-\mathbf{C}\boldsymbol{\Delta}_\xi^{(r)^{-1}}\boldsymbol{\eta}_{\xi y}^{(r)}\big]\\&=\hat{\boldsymbol{\beta}}^r+\boldsymbol{\Delta}_\xi^{(r)^{-1}}\mathbf{C}^\top\big[\mathbf{C}\boldsymbol{\Delta}_\xi^{(r)^{-1}}\mathbf{C}^\top\big]^{-1}\big[\mathbf{d}-\mathbf{C}\hat{\boldsymbol{\beta}}^r\big],\end{aligned}\tag{21}$$

for $r=0,1,\ldots,$ where $\hat{\boldsymbol{\beta}}^{(r)}$, $\boldsymbol{\eta}_{\xi y}^{(r)}$, and $\boldsymbol{\Delta}_\xi^{(r)}$ are obtained from the M-step given in Sec. 3.1. The EM algorithm for estimating the parameters of the model (10) under the restriction $\mathbf{C}\boldsymbol{\beta}=\mathbf{d}$, denoted by $\tilde{\boldsymbol{\theta}}_c$, follows the same procedures given in (15)–(20), replacing $\hat{\boldsymbol{\beta}}$ by $\tilde{\boldsymbol{\beta}}_c$ in the M-step of the algorithm. Note from (21) that the problem of testing linear inequality hypotheses of the form $H_0:\mathbf{C}\boldsymbol{\beta}-\mathbf{d}\geq\mathbf{0}$ can easily be treated using conditions given in Fahrmeir and Klinger (1994) which guarantee that $\tilde{\boldsymbol{\beta}}_c$ corresponds to the inequality restricted estimate. See Cysneiros and Paula (2004) for further details on this approach.

4. The Skew-Normal Linear Mixed Model

In general, a normal linear mixed effects model (N-LMM) is defined as

$$\mathbf{y}_i=\mathbf{X}_i\boldsymbol{\beta}+\mathbf{Z}_i\mathbf{b}_i+\boldsymbol{\epsilon}_i,\quad i=1,\ldots,m,\tag{22}$$

where $\mathbf{y}_i$ is an $(n_i\times 1)$ vector of observed continuous responses for individual i, $\mathbf{X}_i$ of dimension $(n_i\times p)$ 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, $\mathbf{Z}_i$ of dimension $(n_i\times q)$ is the design matrix corresponding to the $(q\times 1)$ random effects vector $\mathbf{b}_i$, and $\boldsymbol{\epsilon}_i$ of dimension $(n_i\times 1)$ is a vector of random errors. It is assumed that the random effects $\mathbf{b}_i$ and the residual components $\boldsymbol{\epsilon}_i$ are independent with $\mathbf{b}_i\overset{\text{iid}}{\sim}N_q(\mathbf{0},\mathbf{D})$ and $\boldsymbol{\epsilon}_i\overset{\text{ind}}{\sim}N_{n_i}(\mathbf{0},\boldsymbol{\Sigma}_i)$. The $q\times q$ covariance matrix $\mathbf{D}$ may be unstructured or structured. The $n_i\times n_i$ covariance matrices $\boldsymbol{\Sigma}_i=\boldsymbol{\Sigma}_i(\gamma)$, $i=1,\ldots,m$, are typically assumed to depend on i through their dimension, being generally parameterized by a parameter vector $\boldsymbol{\gamma}$ of small dimension as, for example, in the constant correlation setting.

Relaxation of the normality assumption on the random effects has been considered by Zhang and Davidian (2001), Verbeke and Lesaffre (1996), Mangder and Zeger (1996), and Tao et al. (1999), among others. Ma et al. (2004) propose a generalized flexible skew-elliptical distribution to represent the random effects density and describe algorithms for ML estimation and Bayesian inference using Markov Chain Monte Carlo methods. Recently, Arellano-Valle et al. (2005) studied a skew-normal linear mixed effects model (SN-LMM) for which the random error and random effects follow skew-normal distributions, and presented EM-type algorithms to perform ML estimation based on the marginal likelihood.

Following Arellano-Valle et al. (2005) (see also Ma et al., 2004), we extend the N-LMM defined above by considering the linear relationship in (22) with the following assumptions:

$$\mathbf{b}_i \overset{\text{iid}}{\sim} SN_q(\mathbf{0}, \mathbf{D}, \boldsymbol{\lambda}_b) \quad \text{and} \quad \boldsymbol{\epsilon}_i \overset{\text{ind}}{\sim} N_{n_i}(\mathbf{0}, \boldsymbol{\Sigma}_i), \tag{23}$$

where $\mathbf{b}_i$ is independent of $\boldsymbol{\epsilon}_i$, $i = 1, \ldots, m$. The asymmetry parameter $\boldsymbol{\lambda}_b$ incorporates asymmetry in the random effects $\mathbf{b}_i$ and consequently in the observed quantities $\mathbf{y}_i$, $i = 1, \ldots, m$. When $\boldsymbol{\lambda}_b = 0$, the model reduces to the usual N-LMM for which inference is treated extensively in the literature.

Note from (3) that the regression set-up defined in (22)–(23) can be written hierarchically as

$$\mathbf{y}_i \mid \mathbf{b}_i \overset{\text{ind}}{\sim} N_{n_i}(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i, \boldsymbol{\Sigma}_i), \tag{24}$$

$$\mathbf{b}_i \mid T_i = t_i \overset{\text{ind}}{\sim} N_q(\mathbf{D}^{1/2}\boldsymbol{\delta}_b t_i, \mathbf{D}^{1/2}(\mathbf{I}_q - \boldsymbol{\delta}_b\boldsymbol{\delta}_b^\top)\mathbf{D}^{1/2}), \tag{25}$$

$$T_i \overset{\text{iid}}{\sim} HN_1(0, 1), \tag{26}$$

$i = 1, \ldots, m$, all independent, where $\boldsymbol{\delta}_b = \boldsymbol{\lambda}_b/(1 + \boldsymbol{\lambda}_b^\top\boldsymbol{\lambda}_b)^{1/2}$.

From Corollary 2 in Arellano-Valle et al. (2005), it also follows that the marginal density of $\mathbf{y}_i$, which is important when conducting inference for the parameter vector $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \boldsymbol{\alpha}^\top, \boldsymbol{\gamma}^\top, \boldsymbol{\lambda}_b^\top)^\top$, where $\boldsymbol{\alpha}$ denotes a minimal set of parameters to determine $\mathbf{D}$ (e.g., the upper triangular elements of $\mathbf{D}$ in the unstructured case), is given by

$$f(\mathbf{y}_i \mid \boldsymbol{\theta}, \boldsymbol{\lambda}_b) = 2\phi_{n_i}(\mathbf{y}_i \mid \mathbf{X}_i\boldsymbol{\beta}, \boldsymbol{\Psi}_i)\Phi_1\big(\bar{\boldsymbol{\lambda}}_{b_i}^\top\boldsymbol{\Psi}_i^{-1/2}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})\big), \tag{27}$$

i.e., $\mathbf{y}_i \overset{\text{ind}}{\sim} SN_{n_i}(\mathbf{X}_i\boldsymbol{\beta}, \boldsymbol{\Psi}_i, \bar{\boldsymbol{\lambda}}_{b_i})$, $i = 1, \ldots, m$, with $\boldsymbol{\Psi}_i = \boldsymbol{\Sigma}_i + \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^T$,

$$\boldsymbol{\Lambda}_i = (\mathbf{D}^{-1} + \mathbf{Z}_i^T\boldsymbol{\Sigma}_i^{-1}\mathbf{Z}_i)^{-1} \quad \text{and} \quad \bar{\boldsymbol{\lambda}}_{b_i} = \frac{\boldsymbol{\Psi}_i^{-1/2}\mathbf{Z}_i\mathbf{D}^{1/2}\boldsymbol{\lambda}_b}{\sqrt{1 + \boldsymbol{\lambda}_b^\top\mathbf{D}^{-1/2}\boldsymbol{\Lambda}_i\mathbf{D}^{-1/2}\boldsymbol{\lambda}_b}}.$$

Hence, the log-likelihood function for $\boldsymbol{\theta}$ given the observed sample $\mathbf{y} = (\mathbf{y}_1^\top, \ldots, \mathbf{y}_n^\top)^\top$ is given by $\ell(\boldsymbol{\theta} \mid \mathbf{y}) = \sum_{i=1}^m \log f(\mathbf{y}_i \mid \boldsymbol{\theta})$.

4.1. *ML Estimation via the EM-Algorithm*

In this section we consider the hierarchical multivariate skew-normal model defined by (24)–(26) with $\mathbf{b}_i$ and the T_i as missing data. Let $\mathbf{y}_c = (\mathbf{y}^\top, \mathbf{b}^\top, \mathbf{t}^\top)^\top$, with $\mathbf{y} = (\mathbf{y}_1^\top, \ldots, \mathbf{y}_n^\top)^\top$, $\mathbf{b} = (\mathbf{b}_1^\top, \ldots, \mathbf{b}_m^\top)^\top$, and $\mathbf{t} = (t_1, \ldots, t_m)^\top$. Thus, under the hierarchical representation (24)–(26), with $\boldsymbol{\Delta}_b = \mathbf{D}^{1/2}\boldsymbol{\delta}_b$ and $\boldsymbol{\Gamma}_b = \mathbf{D} - \boldsymbol{\Delta}_b\boldsymbol{\Delta}_b^\top$, it follows that the complete log-likelihood function is of the form

$$\ell_c(\boldsymbol{\theta}\,|\,\mathbf{y}_c) = \sum_{i=1}^{m}\bigg[-\frac{1}{2}\log|\boldsymbol{\Sigma}_i| - \frac{1}{2}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{Z}_i\mathbf{b}_i)^\top\boldsymbol{\Sigma}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{Z}_i\mathbf{b}_i)$$
$$- \frac{1}{2}\log|\boldsymbol{\Gamma}_b| - \frac{1}{2}(\mathbf{b}_i - \boldsymbol{\Delta}_b t_i)^\top\boldsymbol{\Gamma}_b^{-1}(\mathbf{b}_i - \boldsymbol{\Delta}_b t_i) - \frac{1}{2}t_i^2\bigg] + C, \quad (28)$$

where C is a constant that is independent of the parameter vector $\boldsymbol{\theta}$. As in Sec. 3.1, letting $\hat{\mathbf{b}}_i = \mathrm{E}[\mathbf{b}_i\,|\,\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}, \mathbf{y}_i]$, $\widehat{\boldsymbol{\Omega}}_i = \mathrm{Cov}[\mathbf{b}_i\,|\,\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}, \mathbf{y}_i]$, $\hat{t}_i = \mathrm{E}[T_i\,|\,\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}, \mathbf{y}_i]$, $\hat{t}_i^2 = \mathrm{E}[T_i^2\,|\,\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}, \mathbf{y}_i]$, and $\hat{\mathbf{tb}}_i = \mathrm{E}[T_i\mathbf{b}_i\,|\,\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}, \mathbf{y}_i]$, we obtain after some rather tedious but elementary algebraic manipulation that

$$\hat{t}_i = \hat{\mu}_{Ti} + W_{\Phi_1}\left(\frac{\hat{\mu}_{T_i}}{\widehat{M}_{T_i}}\right)\widehat{M}_{T_i}, \quad \hat{t}_i^2 = \hat{\mu}_{T_i}^2 + \widehat{M}_{T_i}^2 + W_{\Phi_1}\left(\frac{\hat{\mu}_{T_i}}{\widehat{M}_{T_i}}\right)\widehat{M}_{T_i}\hat{\mu}_{T_i},$$
$$\hat{\mathbf{b}}_i = \hat{\mathbf{r}}_i + \hat{\mathbf{s}}_i\,\hat{t}_i, \quad \widehat{\boldsymbol{\Omega}}_i = \widehat{\mathbf{T}}_{b_i}^2 + \hat{\mathbf{s}}_i\hat{\mathbf{s}}_i^\top(\hat{t}_i^2 - (\hat{t}_i)^2), \quad \hat{\mathbf{tb}}_i = \hat{\mathbf{r}}_i\,\hat{t}_i + \hat{\mathbf{s}}_i\,\hat{t}_i^2, \quad (29)$$

where $\widehat{M}_{T_i}^2 = [1 + \widehat{\boldsymbol{\Delta}}_b^\top\mathbf{Z}_i^\top(\widehat{\boldsymbol{\Sigma}}_i + \mathbf{Z}_i\widehat{\boldsymbol{\Gamma}}_b\mathbf{Z}_i^\top)^{-1}\mathbf{Z}_i\widehat{\boldsymbol{\Delta}}_b]^{-1}$, $\hat{\mu}_{T_i} = \widehat{M}_{T_i}^2\widehat{\boldsymbol{\Delta}}_b^\top\mathbf{Z}_i^\top(\widehat{\boldsymbol{\Sigma}}_i + \mathbf{Z}_i\widehat{\boldsymbol{\Gamma}}_b\mathbf{Z}_i^\top)^{-1}(\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}})$, $\widehat{\mathbf{T}}_{b_i}^2 = [\widehat{\boldsymbol{\Gamma}}_b^{-1} + \mathbf{Z}_i^\top\widehat{\boldsymbol{\Sigma}}_i^{-1}\mathbf{Z}_i]^{-1}$, $\hat{\mathbf{r}}_i = \widehat{\mathbf{T}}_{b_i}^2\mathbf{Z}_i^\top\widehat{\boldsymbol{\Sigma}}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}})$, $\hat{\mathbf{s}}_i = (\mathbf{I}_q - \widehat{\mathbf{T}}_{b_i}^2\mathbf{Z}_i^\top\widehat{\boldsymbol{\Sigma}}_i^{-1}\mathbf{Z}_i)\widehat{\boldsymbol{\Delta}}_b$.

It follows that the conditional expectation of the complete log-likelihood function has the form

$$Q(\boldsymbol{\theta}\,|\,\hat{\boldsymbol{\theta}}) = E_{\mathbf{b},\mathbf{T}}[\ell_c(\boldsymbol{\theta}\,|\,\mathbf{y}_c)\,|\,\mathbf{y}, \hat{\boldsymbol{\theta}}] = \sum_{i=1}^{m} Q_i(\boldsymbol{\theta}\,|\,\hat{\boldsymbol{\theta}}), \quad (30)$$

where

$$Q_i(\boldsymbol{\theta}\,|\,\hat{\boldsymbol{\theta}}) = Q_{1i}(\boldsymbol{\beta}, \boldsymbol{\gamma}\,|\,\hat{\boldsymbol{\theta}}) + Q_{2i}(\boldsymbol{\alpha}, \boldsymbol{\lambda}_b\,|\,\hat{\boldsymbol{\theta}}),$$

with

$$Q_{1i}(\boldsymbol{\beta}, \boldsymbol{\gamma}\,|\,\hat{\boldsymbol{\theta}}) = -\frac{1}{2}\log|\boldsymbol{\Sigma}_i| - \frac{1}{2}\big(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{Z}_i\hat{\mathbf{b}}_i\big)^\top\boldsymbol{\Sigma}_i^{-1}(\mathbf{y}_i - \mathbf{x}_i\boldsymbol{\beta} - \mathbf{Z}_j\hat{\mathbf{b}}_i)$$
$$- \frac{1}{2}\mathrm{tr}(\boldsymbol{\Sigma}_i^{-1}\mathbf{Z}_i\widehat{\boldsymbol{\Omega}}_i\mathbf{Z}_i^\top), \quad (31)$$

$$Q_{2i}(\boldsymbol{\alpha}, \boldsymbol{\lambda}_b\,|\,\hat{\boldsymbol{\theta}}) = -\frac{1}{2}\log|\boldsymbol{\Gamma}_b| - \frac{1}{2}\mathrm{tr}\big(\boldsymbol{\Gamma}_b^{-1}(\widehat{\boldsymbol{\Omega}}_i + \hat{\mathbf{b}}_i\hat{\mathbf{b}}_i^\top - 2\widehat{\mathbf{tb}}_i\boldsymbol{\Delta}_b^\top + \hat{t}_i^2\boldsymbol{\Delta}_b\boldsymbol{\Delta}_b^\top)\big), \quad (32)$$

where $\hat{\mathbf{b}}_i$, $\widehat{\boldsymbol{\Omega}}$ and $\widehat{\mathbf{tb}}_i$ are as in (29), $i = 1, \ldots, m$ and $\mathrm{tr}(\mathbf{A})$ indicates the trace of the matrix $\mathbf{A}$.

We then have the following EM algorithm for the SN-LMM defined above:

E-step: Given $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$, compute $\hat{t}_i$, $\hat{t}_i^2$, $\hat{\mathbf{b}}_i$, $\widehat{\boldsymbol{\Omega}}_i$, and $\widehat{\mathbf{tb}}_i$ for $i = 1, \ldots, m$ using (29).

M-step: Update $\hat{\boldsymbol{\theta}}$ by maximizing $Q(\boldsymbol{\theta}\,|\,\hat{\boldsymbol{\theta}})$ over $\boldsymbol{\theta}$, which leads to the following constrained maximization (CM) steps (see Meng and Rubin, 1993):

CM-step 1: Fix $\gamma = \hat{\gamma}$ and update $\boldsymbol{\beta}$ as

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{m} \mathbf{X}_i^\top \widehat{\boldsymbol{\Sigma}}_i^{-1} \mathbf{X}_i\right)^{-1} \sum_{i=1}^{m} \mathbf{X}_i \widehat{\boldsymbol{\Sigma}}_i^{-1} (\mathbf{y}_i - \mathbf{Z}_i \hat{\mathbf{b}}_i). \tag{33}$$

CM-step 2: Fix $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$ and update $\hat{\gamma}$ as

$$\hat{\gamma} = \operatorname{argmax}_{\gamma}\{Q_{1i}(\hat{\boldsymbol{\beta}}, \gamma \mid \hat{\boldsymbol{\theta}})\}. \tag{34}$$

CM-step 3: Update $\widehat{\boldsymbol{\Delta}}_b$ by maximizing $Q(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}})$ over $\boldsymbol{\Delta}_b$, that is,

$$\widehat{\boldsymbol{\Delta}}_b = \frac{\sum_{i=1}^{m} \widehat{\mathbf{tb}}_i}{\sum_{i=1}^{m} \hat{t}_i^2}. \tag{35}$$

CM-step 4: Fix $\boldsymbol{\Delta}_b = \widehat{\boldsymbol{\Delta}}_b$ and update $\widehat{\boldsymbol{\Gamma}}_b$ by maximizing $Q(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}})$ over $\boldsymbol{\Gamma}_b$, which gives,

$$\widehat{\boldsymbol{\Gamma}}_b = \frac{1}{m} \sum_{i=1}^{m} \left(\widehat{\boldsymbol{\Omega}}_i + \hat{\mathbf{b}}_i \hat{\mathbf{b}}_i^\top - 2\widehat{\mathbf{tb}}_i \widehat{\boldsymbol{\Delta}}_b^\top + \hat{t}_i^2 \widehat{\boldsymbol{\Delta}}_b \widehat{\boldsymbol{\Delta}}_b^\top\right). \tag{36}$$

The skewness parameter vector, and the parameters of the scale matrix of the random effects $\mathbf{b}$, can be estimated by noting that $\widehat{\mathbf{D}} = \widehat{\boldsymbol{\Gamma}}_b + \widehat{\boldsymbol{\Delta}}_b \widehat{\boldsymbol{\Delta}}_b^\top$ and $\hat{\boldsymbol{\lambda}}_b = \widehat{\mathbf{D}}^{-1/2} \widehat{\boldsymbol{\Delta}}_b / (1 - \widehat{\boldsymbol{\Delta}}_b^\top \widehat{\mathbf{D}}^{-1} \widehat{\boldsymbol{\Delta}}_b)^{1/2}$ (see Sec. 2.1). Notice that the CM-step 2 only requires a one-parameter search and can be obtained, for instance, using quasi-Newton methods. For the special (and common) situation for which $\boldsymbol{\Sigma}_i = \sigma_e^2 \mathbf{R}_i$, where $\mathbf{R}_i$ is a known matrix of dimension $(n_i \times n_i)$ and $\gamma = \sigma_e^2$, (34) reduces to the following closed form:

$$\hat{\sigma_e^2} = \frac{1}{N} \sum_{i=1}^{m} \left[(\mathbf{y}_i - \mathbf{X}_i \hat{\boldsymbol{\beta}} - \mathbf{Z}_i \hat{\mathbf{b}}_i)^\top \mathbf{R}_i^{-1} (\mathbf{y}_i - \mathbf{X}_i \hat{\boldsymbol{\beta}} - \mathbf{Z}_i \hat{\mathbf{b}}_i) + tr\left(\mathbf{R}_i^{-1} (\mathbf{Z}_i \widehat{\boldsymbol{\Omega}}_i \mathbf{Z}_i^\top)\right)\right].$$

When $\boldsymbol{\lambda}_b = \mathbf{0}$ (or $\boldsymbol{\Delta}_b = \mathbf{0}$), the M-step equations reduce to the equations obtained in Pinheiro and Bates (2000). If $\boldsymbol{\lambda}_b$ is suspected to be close to zero (corresponding to a symmetric model) then it is more appropriate to consider a normal model, since under $\boldsymbol{\lambda}_b = \mathbf{0}$ the information matrix may be singular, as has been proved for simpler models (see Azzalini, 1985; Diciccio and Monti, 2004). Inspection of information criteria such as AIC, BIC, and HQ, can be used in practice to select between the N-LMM (N-IMEM) and SN-LMM (SN-IMEM) fits. See Arellano-Valle et al., 2005 for further details.

4.2. *Restricted Estimation*

Suppose now that we are interested in estimating the parameter vector $\boldsymbol{\beta}$ under k linearly independent restrictions defined as $\mathbf{C}_j^\top \boldsymbol{\beta} - d_j = 0$, where the $\mathbf{C}_j$, $j = 1, \ldots, k$, are $p \times 1$ vectors and d_j, $j = 1, \ldots, k$, are scalars which are assumed to be both known and fixed. The problem is to maximize the complete log-likelihood function $\ell_c(\boldsymbol{\theta} \mid \mathbf{y}_c)$ subject to the linear constraints $\mathbf{C}\boldsymbol{\beta} - \mathbf{d} = \mathbf{0}$, where $\mathbf{C} =$

$(\mathbf{C}_1, \ldots, \mathbf{C}_k)^\top$ and $\mathbf{d} = (d_1, \ldots, d_k)^\top$. Using a similar approach to that employed in Sec. 3.2, one can show that $\boldsymbol{\beta}(\boldsymbol{\Upsilon})$ is the solution of the following iterative process:

$$\begin{aligned}
\widetilde{\boldsymbol{\beta}}_o^{(r+1)} &= \left(\sum_{i=1}^m \mathbf{X}_i^\top \boldsymbol{\Sigma}_i^{-1} \mathbf{X}_i\right)^{-1} \sum_{i=1}^m \mathbf{X}_i^\top \boldsymbol{\Sigma}_i^{-1} (\mathbf{y}_i - \mathbf{Z}_i \hat{\mathbf{b}}_i^{(r)}) + \left(\sum_{i=1}^m \mathbf{X}_i^\top \boldsymbol{\Sigma}_i^{-1} \mathbf{X}_i\right)^{-1} \mathbf{C}^\top \\
&\quad \times \left[\mathbf{C}\left(\sum_{i=1}^m \mathbf{X}_i^\top \boldsymbol{\Sigma}_i^{-1} \mathbf{X}_i\right)^{-1} \mathbf{C}^\top\right]^{-1} \left[\mathbf{d} - \mathbf{C}\left(\sum_{i=1}^m \mathbf{X}_i^\top \boldsymbol{\Sigma}_i^{-1} \mathbf{X}_i\right)^{-1} \sum_{i=1}^m \mathbf{X}_i \boldsymbol{\Sigma}_i^{-1} (\mathbf{y}_i - \mathbf{Z}_i \hat{\mathbf{b}}_i^{(r)})\right] \\
&= \hat{\boldsymbol{\beta}}^{(r)} + \left(\sum_{i=1}^m \mathbf{X}_i^\top \boldsymbol{\Sigma}_i^{-1} \mathbf{X}_i\right)^{-1} \mathbf{C}^\top \left[\mathbf{C}\left(\sum_{i=1}^m \mathbf{X}_i^\top \boldsymbol{\Sigma}_i^{-1} \mathbf{X}_i\right)^{-1} \mathbf{C}^\top\right]^{-1} (\mathbf{d} - \mathbf{C}\hat{\boldsymbol{\beta}}^{(r)}), \qquad (37)
\end{aligned}$$

for $r = 0, 1, \ldots,$ where $\hat{\boldsymbol{\beta}}^{(r)}$ and $\hat{\mathbf{b}}_i^{(r)}$ are obtained from (29) and (33), respectively. The EM algorithm for estimating the parameters of the model (22)–(23) under the restriction $\mathbf{C}\boldsymbol{\beta} = \mathbf{d}$, denoted by $\widetilde{\boldsymbol{\theta}}_o$, follows the same procedures given in Sec. 4.1, replacing $\hat{\boldsymbol{\beta}}$ by $\widetilde{\boldsymbol{\beta}}_o$ in the M-step of the algorithm.

5. Illustrative Examples

In this section, we present results for two real data sets analyzed using the methodology discussed in the previous sections. In both cases we used the convergence criterion

$$\max_{j=1,\ldots,k} \left|(\hat{\theta}_j^{(m+1)} - \hat{\theta}_j^{(m)})/\hat{\theta}_j^{(m)}\right| \leq 10^{-4}$$

for the EM algorithm, were k is the dimension of $\boldsymbol{\theta}$. The starting value for $\boldsymbol{\theta}$ was set equal to the estimates obtained for the normal model, and a range of positive starting values was used for the skewness parameter.

Example 5.1 (Dental Clinical Trial Data Set)**.** In this example, we apply the multivariate null intercept measurement error model developed in Sec. 3 to the longitudinal data set presented in Hadgu and Koch (1999). In that study, 105 adult volunteers with pre-existing dental plaque were assigned at random to two experimental mouth rinses, A ($i = 2, n_2 = 33$) and B ($i = 3, n_3 = 36$), or to a control mouth rinse C ($i = 1, n_1 = 36$). The dental plaque index for each subject was evaluated under these three experimental conditions (mouth rinse), at the start of the trial, after three months and after six months. As the covariate (observed plaque index) is measured imprecisely, one way of analyzing the data is to fit a measurement error model. We do not include intercepts in the model as we consider it reasonable to assume that a zero dental plaque index at the start of the trial should imply a zero expected post-test value; that is, we assume that the dental plaque index should not increase after the use of the mouth rinses. As stated in Sec. 3, the response variable of interest in this context is the true dental plaque index.

The resulting parameter estimates for the SN-IMEM model in (10) and the N-IMEM model are given in Table 1. The notation used is the same as that used in (10), where β_{ij}, $i = 1, 2,$ and $j = 1, 2, 3$ denote the rate of change in dental plaque index with mouth rinses A and B at the start of the trial, after three months and

Table 1
Results from fitting the SN-IMEM and N-IMEM models to the dental plaque index data set. The SE values are estimated asymptotic standard errors

	SN-IMEM		N-IMEM	
Parameter	Estimate	SE	Estimate	SE
β_{11}	0.7020	0.0339	0.7021	0.0340
β_{12}	0.5239	0.0441	0.5241	0.0442
β_{13}	0.5088	0.0317	0.5087	0.0317
β_{21}	0.6857	0.0339	0.6859	0.0340
β_{22}	0.5016	0.0441	0.5017	0.0441
β_{23}	0.4139	0.0317	0.4139	0.0317
$\sigma^2_{\epsilon 1}$	0.2746	0.0460	0.2739	0.0461
$\sigma^2_{\epsilon 2}$	0.4306	0.0752	0.4308	0.0751
$\sigma^2_{\epsilon 3}$	0.2257	0.0377	0.2253	0.0380
σ^2	0.0010	0.0154	0.0021	0.0210
μ_x	2.1082	0.0425	2.5343	0.0325
σ^2_x	0.2907	0.0550	0.1086	0.0210
λ_x	6.1291	6.0782	–	–
log-likelihood	−194.4457		−203.3778	
AIC	1.9757		2.0512	
BIC	2.1400		2.2029	
HQ	2.0422		2.1127	

after six months, respectively. Clearly, β_{ij} values less than 1 indicate dental plaque reduction. The parameter estimates are very similar for both models, except for the estimates of μ_x and σ^2_x. The quoted standard errors were estimated numerically using Matlab's *Jacobian* routine. The AIC, BIC, and HQ values shown at the bottom of Table 1 clearly favor the SN-IMEM over the N-IMEM, supporting the contention of skew departure from normality. This conclusion is also supported by the results from the likelihood ratio test of $H_0: \ \lambda_x = 0$ ($\daleth_{LR} = 17.8642$, p-value $\simeq 0$). It is also corroborated graphically by Fig. 1. Nevertheless, a nominally 95% symmetric confidence interval for λ_x, calculated using the (very large) estimated standard deviation of 6.0782 and large-sample normal approximation, was found to be $(-5.8, 18.0)$, in clear disagreement with the previously quoted results. However, simulation studies conducted by the authors appear to indicate that Wald-type statistics based on the asymptotic covariance matrix, estimated using the observed information matrix, are typically less powerful at detecting skewness than the likelihood ratio statistic.

Considering the hypothesis $H_{01}: \beta_{11} = \beta_{12} = \beta_{13}$, which corresponds to a comparison of the effects, after three months, of the experimental mouth rinses A and B with that of the control mouth rinse C, the restricted ML estimates are given by $\widetilde{\beta}_{11} = \widetilde{\beta}_{12} = \widetilde{\beta}_{13} = 0.5811$, $\widetilde{\beta}_{21} = 0.6857$, $\widetilde{\beta}_{22} = 0.5016$, $\widetilde{\beta}_{23} = 0.4139$, $\widetilde{\sigma}^2_{\epsilon 1} = 0.2746$, $\widetilde{\sigma}^2_{\epsilon 2} = 0.4305$, $\widetilde{\sigma}^2_{\epsilon 3} = 0.2257$, $\widetilde{\sigma}^2 = 0.0011$, $\widetilde{\mu}_x = 2.1082$, $\widetilde{\sigma}^2_x = 0.2907$, and $\widetilde{\lambda}_x = 6.1285$. The H_{01} hypothesis is emphatically rejected since $\daleth_{LR} = 19.59$, corresponding to a p-value of around zero. If we now consider the hypothesis $H_{02}: \beta_{11} = \beta_{21}, \beta_{12} = \beta_{22}$, which corresponds to analyzing whether the control mouth rinse C and the mouth rinse A reduce dental plaque at the same rates over the entire clinical trial, we fail to reject it as $\daleth_{LR} = 0.2418$, corresponding to a p-value in excess of 0.1.

Example 5.2 (Framingham Cholesterol Data Set)**.** To illustrate the use of the EM algorithm for the SN-LMM model described in Sec. 4, we apply it in the analysis of the longitudinal data on cholesterol levels from the famous Framingham heart study. The data set consists of the cholesterol levels over time, age at the start of the study, and gender for $m = 200$ randomly selected individuals. We adopt the same linear mixed model as that used by Arellano-Valle et al. (2005) to analyze these data, namely

$$Y_{ij} = \beta_o + \beta_1 \text{ sex}_i + \beta_2 \text{ age}_i + \beta_3 t_{ij} + b_{oi} + b_{1i} t_{ij} + \epsilon_{ij}\,, \tag{38}$$

where Y_{ij} is the cholesterol level, divided by 100, at the jth time for subject i; t_{ij} is (time − 5)/10, with time measured in years from the start of the study; age_i is age at the start of the study; sex_i is the gender indicator (0 = female, 1 = male). Thus, $\mathbf{x}_{ij} = (1,\ \text{sex}_i, \text{age}_i,\ t_{ij})^\top$, $\mathbf{b}_i = (b_{oi},\ b_{1i})^\top$, and $\mathbf{Z}_{ij} = (1,\ t_{ij})^\top$. The histogram of the cholesterol levels (not shown here) clearly indicates an underlying asymmetric distribution and thus it would seem appropriate to fit an SN-LMM model to the data. Throughout, we assume that $\boldsymbol{\Sigma}_i = \sigma_e^2 \mathbf{I}_{n_i}$, $i = 1, \ldots, 200$.

In an attempt to ensure the global maximum was identified, we used the EM-algorithm with a range of different starting values for $\hat{\boldsymbol{\theta}}_o$ (without any restriction). This approach led to the same estimates on each occasion, a fact that we interpret

Table 2
Results from fitting the SN-LMM and N-LMM models to the Framingham cholesterol data set. d_{11}, d_{12}, and d_{22} are the distinct elements of the matrix $\mathbf{D}^{1/2}$. The SE values are estimated asymptotic standard errors

	SN-LMM		N-LMM	
Parameter	Estimate	SE	Estimate	SE
β_o	1.3555	0.1410	1.5967	0.1543
β_1	−0.0484	0.0491	−0.0631	0.0568
β_2	0.0150	0.0035	0.0184	0.0037
β_3	0.3541	0.0506	0.2817	0.0242
σ_e^2	0.0429	0.0024	0.0434	0.0024
d_{11}	0.5308	0.0489	0.3716	0.0201
d_{12}	0.0016	0.0270	0.0563	0.0179
d_{22}	0.2182	0.0362	0.1868	0.0329
λ_{b1}	12.4602	5.9402	–	
λ_{b2}	−5.9172	2.6839	–	
log-likelihood	−152.0384		−160.9864	
AIC	0.1552		0.1619	
BIC	0.1789		0.1808	
HQ	0.1642		0.1691	

as providing an indication that, at least for these data, the proposed approach is robust with respect to the starting values used.

Table 2 contains the ML estimates, obtained using the EM algorithm, for the parameters of the N-LMM and SN-LMM models, together with their corresponding standard errors calculated numerically using Matlab. The AIC, BIC, and HQ criteria indicate that the SN-LMM presents the best fit, supporting the contention of skew departure from normality. The likelihood ratio test of the hypothesis $H_0: \boldsymbol{\lambda}_b = \mathbf{0}$ leads to the same conclusion ($\daleth_{LR} = 17.8960$, p-value $\simeq 0$). We note that the estimates for $\boldsymbol{\lambda}_b$ given in Table 2 are different from those reported in Arellano-Valle et al. (2005). This is due to the fact that, in their article, these authors computed their estimates assuming that the elements of $\boldsymbol{\lambda}_b$ were positive.

Considering the hypothesis $H_{01}: \beta_2 = 0$, which implies that there is no sex effect on cholesterol level, the restricted ML estimates of the elements of $\boldsymbol{\theta}$ were calculated to be $\widetilde{\beta}_0 = 1.9202$, $\widetilde{\beta}_1 = -0.0531$, $\widetilde{\beta}_3 = 0.3367$, $\widetilde{\sigma}_e^2 = 0.0432$, $\widetilde{d}_{11} = 0.5996$, $\widetilde{d}_{12} = 0.4139$, $\widetilde{d}_{22} = 0.2062$, $\widetilde{\lambda}_{b1} = 11.1831$, $\widetilde{\lambda}_{b2} = -3.5551$. The hypothesis H_{01} is emphatically rejected, since $\daleth_{LR} = 19.98$ and the corresponding p-value is effectively zero.

6. Concluding Remarks

We have developed EM algorithm based methods for estimating the parameters of three multivariate skew-normal regression models. An important characteristic of the results obtained is that closed form expressions were derived for the iterative estimation processes. This was a consequence of the fact that the models possess stochastic representations that can be used to represent them hierarchically. In general, if the underlying distribution is non normal, gains in the precision of the estimates can be achieved (see also Arellano-Valle et al., 2005). We believe that the approaches proposed here can also be extended to the estimation of other multivariate models.

Acknowledgments

We thank the issue guest editor Arthur Pewsey for thoroughly revising the article and pointing out several inconsistencies in the original version. Helpful suggestions from an anonymous referee are also acknowledged. Grants from Fondecyt-1040865 – Chile, CNPq and CAPES – Brazil are also acknowledged.

References

Aoki, R., Bolfarine, H., Achcar J. A., Dorival L. P. (2003). Bayesian analysis of a multivariate null intercept measurement error regression model. *J. Biopharmaceut. Statist.* 13:767–775.

Arellano-Valle, R. B., Azzalini, A. (2006). On the unification of families of skew-normal distributions. *Scand. J. Statist.* In press.

Arellano-Valle, R. B., Bolfarine, H., Lachos, V. H. (2005). Skew-normal linear mixed models. *J. Data Sci.* 3:415–438.

Arellano-Valle, R. B., Genton, M. G. (2005). Fundamental skew distributions. *J. Multivariate Anal.* 96:93–116.

Arellano-Valle, R. B., Ozan, S., Bolfarine, H., Lachos, V. H. (2005). Skew normal measurement error models. *J. Multivariate Anal.* 96:265–281.

Arnold, B. C., Beaver, R. J. (2002). Skewed multivariate models related to hidden truncation and/or selective reporting. *Test* 11:7–54.

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scand. J. Statist.* 12:171–178.

Azzalini, A. (1986). Further results on a class of distributions which includes the normal ones. *Statistica* 46:199–208.

Azzalini, A. (2005). The skew-normal distribution and related multivariate families. *Scand. J. Statist.* 32:159–188.

Azzalini, A., Capitanio, A. (1999). Statistical applications of the multivariate skew normal distribution. *J. Roy. Statist. Soc.* 61:579–602.

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

Cheng, C. L., Van Ness, J. (1999). *Statistical Regression with Measurement Error*. London: Arnold Publishers.

Cysneiros, F. J. A., Paula, G. A. (2004). One-sided tests in linear models with multivariate t-distribution. *Commun. Statist. Simul. Computat.* 33:747–772.

Dempster, A. P., Laird, N. M., Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM-algorithm. *J. Roy. Statist. Soc.* A 39:1–22.

Diciccio, T., Monti, A. C. (2004). Inferential aspects of the skew exponential power distribution. *J. Amer. Statist. Assoc.* 99:439–450.

Fahrmeir, L., Klinger, J. (1994). Estimating and testing generalized linear models under inequality restrictions. *Statist. Pap.* 59:981–995.

Fuller, W. A. (1987). *Measurement Error Models.* New York: John Wiley.

Genton, M. G., ed. (2004). *Skew-Elliptical Distributions and Their Applications: A Journey Beyond Normality*. Boca Raton, FL: Chapman & Hall/CRC.

Genton, M. G., He, L., Liu, X. (2001). Moments of skew-normal random vectors and their quadratic forms. *Statist. Probab. Lett.* 51:319–325.

Genton, M. G., Loperfido, N. (2005). Generalized skew-elliptical distributions and their quadratic forms. *Ann. Instit. Statist. Math.* 57:389–401.

Hadgu, A., Koch, G. (1999). Application of generalized estimating equations to a dental randomized clinical trial. *J. Biopharmaceut. Statist.* 9:161–178.

Hu, P., Tsiatis, A. A., Davidian, M. (1988). Estimating the parameters in the Cox model when covariate variables are measured with errors. *Biometrics* 54:1407–1419.

Johnson, N. L., Kotz, S., Balakrishnan, N. (1994). *Continuous Univariate Distributions*, vol. 1. New York: John Wiley.

Lachos, V. H., Bolfarine, H., Vilca, L. F., Galea, M. (2005). Estimation and influence diagnostic for structural comparative calibration models under the skew-normal distribution. Technical report, RT-MAE 2005-12, IME-USP.

Ma, Y., Genton, M. G., Davidian, M. (2004). Linear mixed models with flexible generalized skew-elliptical random effects. In: Genton, M. G., ed. *Skew-Elliptical Distributions and their Applications: A Journey Beyond Normality*. London: Chapman & Hall/CRC.

Mangder, L. S., Zeger, S. L. (1996). A smooth nonparametric estimate of a mixing distribution using mixtures of Gaussians. *J. Amer. Statist. Assoc.* 91:1141–1151.

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

Nyquist, H. (1991). Restricted estimation of generalized linear models. *Appl. Statist.* 40:133–141.

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

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

Verbeke, G., Lesaffre, E. (1996). A linear mixed-effects model with heterogeneity in the random-effects population. *J. Amer. Statist. Assoc.* 91:217–221.

Verbeke, G., Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. New York: Springer-Verlag.

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