Bayesian measurement error models using finite mixtures of scale mixtures of skew-normal distributions

Celso Rômulo Barbosa Cabral, Nelson Lima de Souza and Jeremias Leão

Department of Statistics, Federal University of Amazonas, Manaus Brazil

ABSTRACT
We present a proposal to deal with the non-normality issue in the context of regression models with measurement errors when both the response and the explanatory variable are observed with error. We extend the normal model by jointly modelling the unobserved covariate and the random errors by a finite mixture of scale mixture of skew-normal distributions. This approach allows us to model data with great flexibility, accommodating skewness, heavy tails, and multi-modality. The main virtue of considering measurement error models under the class of scale mixtures of skew-normal distributions is that they have a nice hierarchical representation which allows easy implementation of inference. In order to illustrate the usefulness of the proposed method some simulation studies are presented and a real dataset (Systemic lupus erythematosus) is analyzed.

ARTICLE HISTORY
Received 20 October 2020
Accepted 13 August 2021

KEYWORDS
Bayesian estimation; finite mixtures; MCMC; skew normal distribution; scale mixtures of skew normal

1. Introduction and motivation

Let us consider the problem of modelling the relationship between two random variables \( y \) and \( x \) through a linear regression model, that is,

\[
y = \alpha + \beta x,
\]

where \( \alpha \) and \( \beta \) are parameters to be estimated. Supposing that these variables are unobservable, we assume that what we actually observe is

\[
X = x + \zeta, \quad \text{and} \quad Y = y + e,
\]

where \( \zeta \) and \( e \) are random errors. This is the so-called measurement error (ME) model. There is a vast literature regarding the inferential aspects of these kinds of models. Comprehensive reviews can be found in Fuller [1], Cheng & Van Ness [2] and Carrol et al. [3].

In general, it is assumed that the variables \( x, \zeta \) and \( e \) are independent and normally distributed. However, there are situations when the true distribution of the latent variable \( x \) departs from normality; that is the case when skewness, outliers and multimodality are present. Then, the choice of more flexible models can be a useful alternative to the normal...
one in order to overcome possible drawbacks. To better understand the phenomena, consider the following description of a real dataset (hereafter the SLE data), which will be used to illustrate the applicability of the methods proposed in this article – see Section 7.

Systemic lupus erythematosus (SLE) is an autoimmune disease that affects many organs and systems. The prevalence and incidence of SLE vary with region, sex, age, ethnicity and time [4]. Clinical manifestations involve skin and joint damages, inflammation of membranes (pleura and pericardium), as well as neurological, hematological and renal alterations. Several studies show that SLE patients with renal disease have high mortality risk [5]. Thus, an important issue is to evaluate the renal function of SLE patients. In order to do so, a prospective study was performed by observing patients with SLE at the Rheumatology Service of the Araújo Lima Outpatient Clinic in Manaus, Brazil [6]. The main goal was to study the relationship between two tests, namely the protein/creatinine ratio taken from an isolated urine sample, and the 24-h proteinuria. The protein/creatinine ratio test is a simple test based on a sample from the first-morning urine. The 24-h proteinuria test is considered a gold standard method, as an early and sensitive marker for the detection of possible renal damage. However, this latter method has some disadvantages. For example, some patients can express annoyance about the need to collect samples for 24 h. The two methods were applied to each of 75 patients of both genders, with 18 years old or more. Besides this, all the patients fit the classification criteria for lupus defined by the American College of Rheumatology (ACR) and the Systemic Lupus International Collaborating Clinics (SLICC). Suppose that $Y$ is the observed protein/creatinine ratio and $X$ is the observed 24-h proteinuria. Figure 1 shows a dispersion plot of $X$ vs. $Y$ (both divided by 1000), where one can clearly note departures from normality. In particular, one can see two distinct subgroups, due to a possible unobserved heterogeneity. In this case, the distribution of the responses is possibly bimodal, and the usual normal regression model cannot be used. Our main goal in this work is to present a model with a flexible distribution for the latent covariate $x$ so as to overcome difficulties like these.

If the distribution of the latent variable $x$ departs from normality, the choice of more flexible models can be a useful alternative to accommodate skewness, heavy tails or multimodality. Some works with this kind of approach are Bolfarine & Arellano-Valle [7], Galea et al. [8], and de Castro & Galea [9]. All these authors consider ME models where the normal assumption for the joint distribution of $x$ and the measurement errors is replaced by the Student-$t$ distribution. The works of Rocha et al. [10] and Matos et al. [11] consider an ME Student-$t$ model where the responses and predictors are both censored.

To allow more flexibility, Lachos et al. [12] and Lachos et al. [13] extended the ME model by considering that the joint distribution of $x$ and the measurement errors belongs to the class of the scale mixtures of skew-normal distributions (hereafter SMSN). The members of this family are extensions of classical symmetric distributions. Thus, we have skew-normal, skew Student-$t$ and skew slash distributions, for example. The extension is obtained by the introduction of a shape parameter, as will be seen in Section 2. A related approach was also used by Tomaya & de Castro [14], by supposing that the distribution of $x$ is skew Student-$t$ and the joint distribution of the errors is Student-$t$ in a heteroscedastic ME model.

When the source of non-normality is the presence of unobserved heterogeneity in the distribution of $x$, an alternative is to model this distribution by a finite mixture of normal densities, as in Carroll et al. [15]. Also, as mentioned in McLachlan & Pell [16], the great flexibility of finite mixtures of normal distributions allows modeling data when there
is the simultaneous occurrence of skewness, discrepant observations, and multimodality. But, even when using normal mixtures, one can have an overestimation of the number of components (that is, the number of densities in the mixture) necessary to capture the asymmetric and/or heavy-tailed nature of each subpopulation [17, sec. 5]. To overcome this problem, Cabral et al. [18] assumed that the latent covariate and the random observational errors are jointly modelled by a finite mixture of skew Student-$t$ distributions, developing an EM-type algorithm for inference. Here, we extend this work, by supposing that this joint distribution is a finite mixture with components in the SMSN family. Also, a Bayesian inferential approach is adopted. Bayesian methods can be more suitable in this case, because for our real dataset the sample size is moderate compared with the number of parameters. This fact is, probably, an inferential issue when proceeding maximum likelihood estimation.

The remainder of the paper is organized as follows. In Section 2, for the sake of completeness, we review the basic concepts of the skew-normal distribution and its scale mixtures and also explore the concept of finite mixtures of distributions in this family. In Section 3 we define our extension of the normal measurement error model. In Section 4, we develop the MCMC-type algorithm for Bayesian inference. In Section 5, we discuss model selection and in Sections 6 and 7, we give numerical examples using both simulated and real data to illustrate the performance of the proposed method.

2. The skew normal distribution, the SMSN family, and mixtures

The concept of skew normal (SN) distribution has a long history in the probability and statistical literature, going back to works like Birnbaum [19] and Nelson [20]. However, there is no doubt that the most popular approach is that given by Adelchi Azzalini and colleagues when they presented extensions for the univariate and multivariate normal
distributions in seminal papers like Azzalini [21] and Azzalini & Dalla Valle [22], respectively, followed by unification efforts in Arellano-Valle & Azzalini [23] and Azzalini & Capitanio [24]. Here we use a definition of skew normal distribution as a member of the fundamental skew-normal distribution family (FUSN), presented in Arellano-Valle & Genton [25].

In what follows \( N_q(\mu, \Omega) \) denotes the \( q \)-variate normal distribution with mean vector \( \mu \) and covariance matrix \( \Sigma \), \( N_q(\cdot|\mu, \Omega) \) is the respective probability density function (pdf) and \( \Phi(\cdot) \) is the standard normal distribution function.

**Definition 2.1:** Let \( X_0 \sim N(0,1) \) and \( X_1 \sim N_q(\mu, \Sigma) \) be independent, where \( \Sigma \) is positive definite. Let \( \Delta \) be a \( q \)-dimensional vector. We say that the distribution of \( X = \Delta|X_0| + X_1 \),

is skew normal with location vector \( \mu \), scale matrix \( \Sigma \) and shape vector \( \Delta \). We use the notation \( X \sim SN_q(\mu, \Sigma, \Delta) \).

The distribution of \( X \) belongs to the FUSN family because it is the distribution of \( V|X_0 > 0 \), where \( V = \Delta X_0 + X_1 \). Since \( V \) and \( X_0 \) have jointly a normal distribution, it is straightforward to prove that \( X \) has pdf given by

\[
SN_q(x|\mu, \Sigma, \Delta) = 2N_q(x|\mu, \Omega)\Phi(\lambda^T(y - \mu)), \tag{1}
\]

where

\[
\Omega = \Sigma + \Delta\Delta^T \quad \text{and} \quad \lambda = \frac{\Omega^{-1}\Delta}{(1 - \Delta^T\Omega^{-1}\Delta)^{1/2}}. \tag{2}
\]

Also, we recover \( \Delta \) and \( \Sigma \) by using

\[
\Delta = \Omega \delta, \quad \Sigma = \Omega - \Omega \delta \delta^T \Omega,
\]

where \( \delta = \lambda/(1 + \lambda^T\Omega\lambda)^{1/2} \). Notice that the case \( \lambda = 0 \) (equivalently \( \Delta = 0 \)) corresponds to the usual \( q \)-variate normal distribution. Also, the SN given in Definition 2.1 is the same used before in works like Lachos et al. [12] and Cabral et al. [18], defined by its pdf as \( f(x) = 2N_q(x|\mu, \Omega)\Phi(\lambda^T\Omega^{-1/2}(x - \mu)) \), where \( \Omega^{-1/2} \) is the inverse of the square root of \( \Omega \). Equation (1) is obtained through the parameterization \( \lambda = \Omega^{-1/2}\lambda^* \).

**Definition 2.2:** We say that the distribution of the \( q \)-dimensional random vector \( Y \) belongs to the family of scale mixtures of skew normal (SMSN) distributions when

\[
Y = \mu + U^{-1/2}X, \tag{3}
\]

where \( \mu \) is a \( q \)-dimensional vector of constants, \( X \sim SN_q(0, \Sigma, \Delta) \) and \( U \) is a positive random variable independent of \( X \) having distribution function \( H(\cdot|\nu) \).

Here \( \nu \) is a (possibly multivariate) parameter indexing the distribution of \( U \), which is known as the scale factor. \( H(\cdot|\nu) \) is called the mixing distribution function. We write \( Y \sim \)
SMSN_q(\(\mu, \Sigma, \Delta, v\)). By Definitions 2.1 and (2.2),
\[
Y|U = u \sim SN_q(\mu, u^{-1}\Sigma, u^{-1/2}\Delta),
\]
which implies that the marginal pdf of \(Y\) is
\[
SMSN_q(y|\mu, \Sigma, \Delta, v) = 2 \int_0^\infty N_q(y|\mu, u^{-1}\Omega) \Phi(u^{1/2}\lambda^\top(y - \mu)) \, dH(u|v),
\]
where \(\Omega\) and \(\lambda\) are given in (2).

Depending on the distribution of the scale factor \(U\) we have a different member of the SMSN family. For example, if \(P(U = 1) = 1\) we have the skew normal distribution; \(U \sim \text{Gamma}(\nu/2, \nu/2)\), with \(\nu > 0\), corresponds to the skew Student-t distribution – here we denote by \(\text{Gamma}(a,b)\) the gamma distribution with mean \(a/b\) and variance \(a/b^2\), with \(a, b > 0\); \(U \sim \text{Beta}(\nu, 1)\), with pdf \(f(u|\nu) = \nu u^{\nu-1}, 0 < u < 1, \nu > 0\), corresponds to the skew slash distribution; If \(U\) is binary with \(P(U = \tau) = \rho = 1 - P(U = 1)\), where \(0 < \tau, \rho < 1\) (and therefore \(v = (\tau, \rho)^\top\)), we have the skew contaminated normal distribution. Obviously, there are other distributions in the SMSN family, but for illustrative purposes, we restrict ourselves to these. The SMSN family, first defined by Branco & Dey [26], includes the class of the scale mixtures of normal (SMN) distributions, defined by Andrews & Mallows [27] where normality is assumed for \(X\) in (3) (and so \(\Delta = 0\)). In this case, we use the notations \(Y \sim \text{SMN}_q(\mu, \Sigma, v)\) and \(\text{SMN}_q(\cdot|\mu, \Sigma, v)\) for the respective pdf. Obviously, this class contains the normal, Student-t, slash and contaminated normal distributions.

The skew Student-t pdf is given by:
\[
ST_q(y|\mu, \Sigma, \Delta, v) = 2t_q(y|\mu, \Omega, v) \left[ \left( \frac{v + p}{v + (y - \mu)^\top \Omega^{-1}(y - \mu)} \right)^{1/2} \lambda^\top(y - \mu)|v + p \right],
\]
where \(t_q(\cdot|\mu, \Omega, v)\) and \(T(\cdot|v + p)\) denote, respectively, the pdf of the \(q\)-variate Student-t distribution with location vector \(\mu\), scale matrix \(\Omega\) and \(v\) degrees of freedom, and the distribution function of the standard univariate Student-t distribution with \(v + p\) degrees of freedom, and \(\Omega\) and \(\lambda\) are given in (2) – for a proof, see Branco & Dey [26].

The skew slash distribution has pdf
\[
SSL_q(y|\mu, \Sigma, \Delta, v) = 2v \int_0^1 u^{v-1}N_p(y|\mu, u^{-1}\Omega) \Phi(u^{1/2}\lambda^\top(y - \mu)) \, du,
\]
which can be evaluated using the \(\mathbb{R}\) function \texttt{integrate} [28], for example.

The skew-contaminated normal distribution has pdf
\[
SCN_q(y|\mu, \Sigma, \Delta, v) = 2 \left\{ \rho N_q(y|\mu, \tau^{-1}\Omega) \Phi(\tau^{1/2}\lambda^\top(y - \mu)) + (1 - \rho)N_q(y|\mu, \Omega) \Phi(\lambda^\top(y - \mu)) \right\},
\]
which comes directly from the definition.
From Definitions 2.1 and 2.2, we have that affine transforms of a SMSN distribution are still SMSN. That is, if \( C \) is an \( m \times q \) matrix with rank \( m \), \( d \) is an \( m \)-dimensional vector and \( Y \sim \text{SMSN}_q(\mu_j, \Sigma_j, \Delta, v_j) \), then \( CY + d \sim \text{SMSN}_m(C\mu + d, C\Sigma C^\top, C\Delta, v) \).

A finite mixture of SMSN distributions with \( G \) components is defined by its pdf as

\[
g(y|\Theta) = \sum_{j=1}^{G} p_j \text{SMSN}_q(y|\theta_j),
\]

where \( p_j \geq 0 \) are such that \( \sum_j^G p_j = 1 \), \( \theta_j = (\mu_j, \Sigma_j, \Delta_j, v_j) \) and \( \Theta = (\theta_1, \ldots, \theta_G, p_1, \ldots, p_G) \). The pdf \( \text{SMSN}(\cdot|\theta_j) \) is named the \( j \)th mixture component and \( p_j \) is the corresponding weight. Hereafter, we call (4) as the FMSMSN model. A hierarchical representation of this model is given by \( Y|S = j \sim \text{SMSN}_q(y|\theta_j) \), where \( S \) is a discrete latent variable with probability function \( P(S = j) = p_j, j = 1, \ldots, G \). It is interpreted as a classification variable: given that \( S = j \) then we know that the underlying subject came from a population with distribution \( \text{SMSN}(\cdot|\theta_j) \). Then, using Definitions 2.1 and 2.2, we have the following hierarchical representation for \( Y \) distributed as FMSMSN:

\[
Y|S = j, U = u, T = t \sim \text{N}_q(\mu_j + \Delta_j t, u^{-1} \Sigma_j);
\]

(5)

\[
T|U = u \sim \text{TN}(0, u^{-1}, (0, \infty));
\]

(6)

\[
U \sim H(\cdot|v_j)
\]

(7)

\[
P(S = j) = p_j, \quad j = 1, \ldots, G,
\]

(8)

with \( \text{TN}(\mu, \sigma^2, A) \) denoting a truncated normal distribution, which is the distribution of \( W|W \in A \), where \( W \sim \text{N}(\mu, \sigma^2) \). This representation is useful to obtain a MCMC-type algorithm to perform posterior inference for the proposed model that will be presented next, and also to generate artificial samples from an FMSMSN distribution. For more details about FMSMSN distributions, see Dávila et al. [29].

3. The SMSN mixture measurement error model

The ME model can be put in a more general setting, by considering a multivariate unobserved response \( y = (y_1, \ldots, y_r)^\top \). Thus, we intend to model the relationship between \( y \) and \( x \) by assuming that

\[
y = \alpha + \beta x,
\]

where \( \alpha \) and \( \beta \) are \( r \)-dimensional vectors of unknown regression parameters. Let \( p = r + 1 \) and suppose that \( y \) and \( x \) are observed with error. What we actually observe is the \( p \)-dimensional random vector \( Z = (X, Y^\top)^\top \), such that \( X = x + \zeta \) and \( Y = y + \varepsilon \). Thus,

\[
Z = a + bx + \varepsilon,
\]

(9)

where \( a = (0, \alpha^\top)^\top, b = (1, \beta^\top)^\top, \varepsilon = (\zeta, \varepsilon^\top)^\top, \zeta \) and \( \varepsilon \) are errors when observing \( x \) and \( y \), respectively. Alternatively, defining \( R = (x, \varepsilon^\top)^\top \), the ME model can be written as:

\[
Z = a + BR,
\]

(10)

where \( B = [b I_p] \) is a \( p \times (p + 1) \) partitioned matrix with first column equal to \( b \) and \( I_p \) is the \( p \times p \) identity matrix. In general, it is supposed that \( x, \zeta \) and \( \varepsilon \) are independent, with
where, like in (8), $W$ we propose to extend their model, by supposing that the assumption is not so restrictive; see, for example, Cabral et al. [31], where the linear mixed measurement error model is equivalent to the following stochastic representation, see Equation (9):

$$x \sim \text{SMSN}((\mu_j, \gamma_j^2, \Delta_j, \nu), j = 1, \ldots, G),$$

where, like in (8), $P(S = j) = p_j$. Thus, the pdf of $x$ is the mixture $\sum_{j=1}^{G} p_j \text{SMSN}(\cdot; \mu_j, \gamma_j^2, \Delta_j, \nu)$. We can write the assumptions above as:

$$R|S = j \sim \text{SMSN}_1 + p \left( (\mu_j, 0_p^\top)^\top, \text{block diag}\{\gamma_j^2, \Omega_j\}, (\Delta_j, 0_p^\top)^\top, \nu \right),$$

which implies that the marginal distribution of $R$ is also FMSMSN. Note that we are supposing that the scale factor parameter $\nu$ is the same for all components of the mixture. This assumption is not so restrictive; see, for example, Cabral et al. [31], where the linear mixed model, which has a similar structure, is investigated. We call this model the SMSN finite mixture measurement error model, which will be denoted by FMSMSN-ME or FMSN-ME, FMST-ME, etc. if we use the specific distributions of the family. From (5), we have that

$$R|(S = j, U = u, T = t) \sim N_{1+p} \left( (\mu_j + \Delta_j t, 0_p^\top)^\top, u^{-1} \text{block diag}\{\gamma_j^2, \Omega_j\} \right),$$

where the distributions of $(T, U)$ and $S$ are given by (6)–(7) and (8), respectively. As $\epsilon|U = u \sim N_p(0, u^{-1}\Omega)$, the distribution of the vector of observations $Z = (X, Y^\top)^\top$ has the following stochastic representation, see Equation (9):

$$Z|(x, U = u) \sim N_p(a + bx, u^{-1}\Omega);$$

$$x|(S = j, U = u, T = t) \sim N(\mu_j + \Delta_j t, u^{-1}\gamma_j^2),$$

$$T|U = u \sim T(0, u^{-1}, (0, \infty));$$

$$U \sim H(\cdot|\nu);$$

$$P(S = j) = p_j, \quad j = 1, \ldots, G. \quad (13)$$

An alternative representation can be obtained by integrating out the latent variable $x$. From Equations (10) and (12), the first two equations of representation (13) can be replaced by

$$Z|(S = j, U = u, T = t) \sim N_p \left( a + \mu_j b + \Delta_j b t, u^{-1}(\gamma_j^2 b b^\top + \Omega) \right).$$

These representations are useful to obtain a MCMC-type algorithm to perform posterior inference and also to simulate samples from the FMSMSN-ME model.

Let $\Theta = (\alpha^\top, \beta^\top, \mu^\top, \Delta^\top, \gamma^\top, \omega^\top, p^\top, v^\top)^\top$ be the vector of parameters to be estimated, where $\mu = (\mu_1, \ldots, \mu_G)^\top, \Delta = (\Delta_1, \ldots, \Delta_G)^\top, \gamma = (\gamma_1^2, \ldots, \gamma_G^2)^\top, \omega = \ldots$. \(\square\)
\((\omega_0^2, \omega_1^2, \ldots, \omega_r^2)^\top\) and \(p = (p_1, \ldots, p_G)^\top\). Denoting the conditional pdf of \(Z|\Theta\) by \(\pi(z|\Theta)\), Equations (5)–(8) and (14) imply that

\[
\pi(z|\Theta) = \sum_{j=1}^{G} p_j \text{SMSN}_p(z|\xi_j, \Sigma_j, \Lambda_j, \nu),
\]

where

\[
\xi_j = a + \mu_j b, \quad \Lambda_j = \Delta_j b, \quad \text{and}
\]

\[
\Sigma_j = \gamma_j^2 b b^\top + \Omega = \begin{pmatrix}
\gamma_j^2 + \omega_0^2 & \gamma_j^2 \beta^\top \\
\gamma_j^2 \beta & \gamma_j^2 \beta^\top + \Omega_c
\end{pmatrix}.
\]

### 4. Posterior inference

Let \(z_1, \ldots, z_n\) be an observed random sample from the FMSMSN-ME model. The likelihood function is given by \(\prod_{i=1}^{n} \pi(z_i|\Theta)\), where \(\pi(\cdot|\Theta)\) is given in Equation (15). The prior specification for each of the parameters \(\alpha, \beta, \mu\) and \(\Delta\) is multivariate normal. Regarding the dispersion parameters in \(\gamma\), we adopt the hierarchical prior defined as

\[
\gamma_j^{-2} f \sim \text{Gamma}(e, f) \quad j = 1, \ldots, G; \quad f \sim \text{Gamma}(g, h).
\]

This hierarchical prior setup follows Richardson & Green [32], where the univariate normal mixture case is investigated. Also, we fix \(\omega_i^{-2} \sim \text{Gamma}(l, m), i = 0, 1, \ldots, r\). For the vector of weights, we apply the usual assumption \(p \sim \text{Dir}(k_1, \ldots, k_G)\), that is, a Dirichlet distribution with known positive hyperparameters. In all applications presented in this text, we have chosen hyperparameter values of the prior distributions that express little prior knowledge. Thus, the prior covariance matrices of \(\alpha, \beta, \mu\) and \(\Delta\) are assumed to be diagonal with large variances, the hyperparameters \(g, h, l\) and \(m\) are small and positive (in general we fix the hyperparameter values of the gamma priors equal to 0.01) and \(\kappa_1 = \cdots = \kappa_G = 1\).

Each specific model in the SMSN class has a scale factor parameter \(\nu\) with specific interpretation, deserving a different treatment for prior choice. For instance, there are several suggestions for estimating the unknown degrees of freedom of the Student-\(t\) model; see the discussions in Fonseca et al. [33] and Garay et al. [34]. Here, we do not treat this issue in depth but adopt prior choices that have been useful for our purposes. For example, for the FMST-ME model, we fix as prior for \(\nu\) an exponential distribution with parameter \(\lambda > 0\) with a second level of hierarchy given by \(\lambda \sim \text{U}(\lambda_0, \lambda_1)\) (a uniform distribution on the interval \((\lambda_0, \lambda_1)\)), where \(0 < \lambda_0 < \lambda_1\). In general, we adopt \(\lambda_0 = 0.04\) and \(\lambda_1 = 0.5\). See Congdon [35, p. 171] for more details. For the FMSSSL-ME model, \(\nu \sim \text{Gamma}(\phi_{sl}, \psi_{sl})\), where \(\phi_{sl}\) and \(\psi_{sl}\) are small and positive. For the FMSCN-ME model, a simple prior setup can be considered as \(\rho \sim \text{beta}(\rho_0, \rho_1)\) and \(\tau \sim \text{beta}(\tau_0, \tau_1)\), where \(\rho_0, \rho_1, \tau_0\) and \(\tau_1\) are positive. In general, we adopt a uniform distribution as a prior for these parameters, that is, \(\rho_0 = \rho_1 = \tau_0 = \tau_1 = 1\). Assuming prior independence, the posterior distribution is given
by

\[
\pi(\Theta|z_1, \ldots, z_n) \propto \left( \prod_{i=1}^{n} \pi(z_i|\Theta) \right) \pi(\alpha)\pi(\beta)\pi(\mu)\pi(\Delta)\pi(p)\pi(v|\lambda)\pi(\lambda) \times \left( \prod_{j=0}^{r} \omega_j^2 \right) \left( \prod_{j=1}^{G} \pi(y_j^2|f) \right) \pi(f).
\]

(17)

The hyperparameter \(\lambda\) must be dropped in the cases of the FMSSL-ME and FMSCN-ME models.

Because of the nonstandard form of the posterior distribution (17), the computation of posterior moments estimates is a very hard task. Also, it is not easy to generate samples from this posterior using traditional Monte Carlo methods. A reliable alternative is to develop a MCMC-type algorithm. Using existing Bayesian software like JAGS [36] or Stan [37], this can be easily implemented through representations (13) and (14) – the last one is our choice to carry out the computations. Alternatively, these representations are useful to develop a Gibbs-type algorithm to be implemented using existing software like R [28].

5. Model selection

The deviance information criterion (DIC) [38] is a common Bayesian tool to compare a given set of candidate models. For a given sample \(z = (z_1^T, \ldots, z_n^T)^T\) let

\[
D(z, \Theta) = -2 \sum_{i=1}^{n} \log \pi(z_i|\Theta)
\]

be the deviance – for the FMSMSN-ME model the expression for \(\pi(\cdot|\Theta)\) is given in Equation (15). In this case, in order to simplify the notation, we write \(D(\Theta)\). Let

\[
\overline{D}(\Theta) = E[D(\Theta)|z] = -2 \sum_{i=1}^{n} E[\log \pi(z_i|\Theta)|z]
\]

be the posterior mean deviance. The measure

\[
\tau_D = \overline{D}(\Theta) - D(\widetilde{\Theta}),
\]

(19)

where \(\widetilde{\Theta}\) is an estimator of \(\Theta\), is called the effective dimension (usually, \(\widetilde{\Theta}\) is the posterior mean \(\overline{\Theta} = E[\Theta|z]\)). The DIC is defined as \(\text{DIC} = D(\overline{\Theta}) + 2\tau_D\), having the following expression

\[
\text{DIC} = -4 \sum_{i=1}^{n} E[\log \pi(z_i|\Theta)|z] + 2 \sum_{i=1}^{n} \log \pi(z_i|\overline{\Theta}).
\]

(20)

The terms \(D(\overline{\Theta})\) and \(2\tau_D\) are interpreted as a measure of fit and a penalty for model complexity, respectively. It is well known that there are some issues with this definition. For example, \(\tau_D\) is not invariant to reparameterizations. That is, different parameterizations
can produce different values of $\tau_D$, and hence different values of DIC. Also, in the mixture model case the posterior mean $\overline{\Theta}$ can be a poor choice for $\Theta$, mainly because the finite mixture likelihood is invariant under permutations of the component labels (this property of the likelihood is usually called label switching). If the prior is also invariant with respect to the labels, all posterior means will be equal, and the plug-in mixture $\pi(z_i|\Theta)$ will have only one component. As a consequence, the estimator $D(\overline{\Theta})$ of $D(\Theta)$ is unreasonable, and expression (20) is useless. For more details, see the discussion in Stephens [39, p. 13].

In the context of finite mixture models, a more applicable definition of DIC can be found in Celeux et al. [40], see also Spiegelhalter et al. [41]. Observe that, while the statistics $\pi(z_i|\Theta)$ is affected by label switching, the posterior predictive density evaluated at $z_i$, given by $E[\pi(z_i|\Theta)|z]$, is not. Thus, it is more reasonable to consider the latter as an estimator of $\pi(z_i|\Theta)$ in expression (18). Then, instead of $D(\overline{\Theta})$, we use $-2 \sum_{i=1}^{n} \log E[\pi(z_i|\Theta)|z]$ as an estimator of $D(\Theta)$ in expression (19), resulting in the following alternative definition of DIC:

$$DIC = -4 \sum_{i=1}^{n} E[\log \pi(z_i|\Theta)|z] + 2 \sum_{i=1}^{n} \log E[\pi(z_i|\Theta)|z].$$

Also, defining DIC in this way provides invariance to reparameterization.

In general, it is a hard task to obtain closed-form expressions for the posterior mean $E[\log \pi(z_i|\Theta)|z]$ and for the posterior predictive density $E[\pi(z_i|\Theta)|z]$, but these integrals can be easily approximated using posterior MCMC samples. Let $\Theta^{(l)}$ be the MCMC sample generated at the $l$th step of the algorithm, $l = 1, \ldots, L$. Then, we have the following approximation for the DIC:

$$\frac{-4}{L} \sum_{l=1}^{L} \sum_{i=1}^{n} \log \pi(z_i|\Theta^{(l)}) + 2 \sum_{i=1}^{n} \log \left( \frac{1}{L} \sum_{l=1}^{L} \pi(z_i|\Theta^{(l)}) \right).$$

6. Simulation studies

We present simulation studies in order to show the applicability of our proposed method.

6.1. Simulation study 1 – parameter recovery

The aim of this study is to analyse the performance of the proposed method by studying some frequentist properties of the estimates. In order to do so, an experiment was carried out as follows. Fixing $r = 2$ and $G = 2$, we first generated 100 datasets of size $n$ from the FMST-ME model with the following parameter setup: $\alpha = (0.4, 0.1)^T$, $\beta = (0.8, 0.9)^T$, $\omega_0^2 = 0.2$, $\omega_1^2 = 0.3$, $\omega_2^2 = 0.4$, $\mu_1 = 2$, $\mu_2 = 8$, $\Delta_1 = -2$, $\Delta_2 = 2$, $\gamma_1^2 = \gamma_2^2 = 0.1$, $\rho_1 = 0.7$ and $\nu = 3$.

For each dataset and for each parameter, we obtained an approximation of the posterior mean estimate through MCMC samples. For this purpose, we drew 25,000 MCMC posterior samples with a burn-in of 5,000 iterations and thinning of 30 iterations. We considered the sample sizes $n = 50, 100, 500$. Then, the experiment was repeated for the FMSN-ME, FMSSL-ME and FMSCN-ME models with the same parameter setup, except for the FMSCN-ME model, in which case we fixed $\rho = 0.7$ and $\tau = 0.3$. The average and
are presented in Table 1, where PV is the parameter value used to generate the dataset. The results are very satisfactory, even for the relatively small sample size $n = 50$. This can be confirmed by inspecting some adjusted boxplots in Figure 2. To save space we only exhibit the boxplots for the FMST-ME case, excluding the parameters of the mixture.

### 6.2. Simulation study 2 – the flexibility of the FMSMSN-ME model

We now study the performance of the proposed model when fitting data generated from a measurement error model with a latent covariate having a distribution that is a finite mixture of normal inverse Gaussian distributions (NIG). This experiment is similar to that carried out by Cabral et al. [18]; more details can be found in this reference. The main

| Parameters | FMSN-ME | FMSCN-ME |
|------------|---------|----------|
|            | $n = 50$ | $n = 100$ | $n = 500$ | $n = 50$ | $n = 100$ | $n = 500$ |
| $\Delta_1$ | $-2.0$  | $-2.0$ | $-2.0$ | $-2.0$ | $-2.0$ | $-2.0$ |
| $\Delta_2$ | 2.0     | 1.587   | 1.877   | 2.041   | 1.141   | 2.108   | 2.143   |
| $\omega_1$ | 0.2     | 0.207   | 0.213   | 0.205   | 0.196   | 0.217   | 0.226   |
| $\omega_2$ | 0.3     | 0.309   | 0.313   | 0.301   | 0.337   | 0.372   | 0.338   |
| $\omega_3$ | 0.4     | 0.433   | 0.419   | 0.410   | 0.496   | 0.509   | 0.452   |
| $\alpha_1$ | 0.4     | 0.391   | 0.402   | 0.390   | 0.415   | 0.397   | 0.401   |
| $\alpha_2$ | 0.1     | 0.102   | 0.100   | 0.100   | 0.123   | 0.114   | 0.105   |
| $\beta_1$  | 0.8     | 0.802   | 0.802   | 0.800   | 0.795   | 0.799   | 0.799   |
| $\beta_2$  | 0.9     | 0.897   | 0.902   | 0.900   | 0.892   | 0.900   | 0.899   |
| $\gamma_1$ | 0.1     | 0.207   | 0.122   | 0.100   | 0.216   | 0.132   | 0.084   |
| $\gamma_2$ | 0.1     | 0.222   | 0.140   | 0.100   | 0.258   | 0.155   | 0.097   |
| $\mu_1$    | 2.0     | 1.787   | 1.992   | 2.003   | 1.795   | 2.042   | 2.038   |
| $\mu_2$    | 8.0     | 8.394   | 8.121   | 7.987   | 8.939   | 8.099   | 7.973   |
| $\nu_1$    | 0.7     | 0.694   | 0.701   | 0.700   | 0.680   | 0.692   | 0.697   |
| $\nu_2$    | 0.3     | 0.305   | 0.298   | 0.299   | 0.319   | 0.307   | 0.303   |
| $\rho$     | 0.7     | –       | –       | –       | 0.668   | 0.647   | 0.654   |
| $\tau$     | 0.3     | –       | –       | –       | 0.320   | 0.354   | 0.326   |
| $\nu$      | 3.0     | –       | –       | –       | –       | –       | –       |

| Parameters | FMSSL-ME | FMSCN-ME |
|------------|----------|----------|
|            | $n = 50$ | $n = 100$ | $n = 500$ | $n = 50$ | $n = 100$ | $n = 500$ |
| $\Delta_1$ | $-2.0$  | $-2.100$ | $-2.066$ | $-2.058$ | $-2.078$ | $-2.045$ |
| $\Delta_2$ | 2.0     | 1.461   | 1.988   | 2.029   | 1.903   | 2.079   | 2.026   |
| $\Omega_1$ | 0.2     | 0.233   | 0.228   | 0.211   | 0.236   | 0.220   | 0.207   |
| $\Omega_2$ | 0.3     | 0.334   | 0.343   | 0.310   | 0.342   | 0.325   | 0.303   |
| $\Omega_3$ | 0.4     | 0.453   | 0.451   | 0.415   | 0.476   | 0.429   | 0.411   |
| $\alpha_1$ | 0.4     | 0.389   | 0.394   | 0.396   | 0.397   | 0.401   | 0.392   |
| $\alpha_2$ | 0.1     | 0.109   | 0.118   | 0.106   | 0.093   | 0.099   | 0.088   |
| $\beta_1$  | 0.8     | 0.802   | 0.798   | 0.801   | 0.796   | 0.801   | 0.800   |
| $\beta_2$  | 0.9     | 0.899   | 0.897   | 0.900   | 0.899   | 0.900   | 0.901   |
| $\gamma_1^2$ | 0.1    | 0.117   | 0.102   | 0.081   | 0.227   | 0.161   | 0.092   |
| $\gamma_2^2$ | 0.1    | 0.144   | 0.119   | 0.094   | 0.272   | 0.179   | 0.104   |
| $\mu_1$    | 2.0     | 1.845   | 2.019   | 2.033   | 1.939   | 2.018   | 2.029   |
| $\mu_2$    | 8.0     | 8.580   | 8.097   | 7.978   | 8.249   | 8.041   | 7.982   |
| $p_1$      | 0.7     | 0.688   | 0.687   | 0.699   | 0.671   | 0.701   | 0.700   |
| $p_2$      | 0.3     | 0.311   | 0.312   | 0.301   | 0.328   | 0.298   | 0.299   |
| $\nu$      | 3.0     | 6.308   | 6.053   | 3.554   | 3.855   | 3.478   | 3.111   |

| Table 1. Simulation study 1: average and standard deviation values (in parentheses) computed across 100 posterior mean estimates of the parameters in the FMSMSN-ME model. |
motivation is that the NIG distribution, contrary to SMSN class, is not a sub-family of the class of skew-elliptical distributions.

The NIG distribution is a scale mixture of a normal distribution and an inverse Gaussian (IG) distribution. We say that a random variable $U$ has an IG distribution when its density is given by

$$g(u) = \frac{\delta}{\sqrt{2\pi}} u^{-3/2} \exp \left\{ -\frac{1}{2} \left( \frac{\delta^2}{u} + \gamma^2 u - 2\delta \gamma \right) \right\}, \quad u > 0,$$

where $\gamma > 0$ and $\delta > 0$. In this case, we use the notation $U \sim \text{IG}(\gamma, \delta)$.

**Definition 6.1:** We say that the random vector $X$ has a $p-$dimensional NIG distribution if it admits the representation

$$X | U = u \sim N_p(\mu + u\Delta \lambda, u\Delta), \quad U \sim \text{IG}(\gamma, \delta),$$

where $\mu$ and $\lambda$ are $p$-dimensional vectors of parameters, $\Delta$ is a $p \times p$ positive definite matrix of parameters and $\gamma$ and $\delta$ are positive parameters.

We use the notation $X \sim \text{NIG}(\mu, \Delta, \lambda, \gamma, \delta)$. It is assumed that $\det(\Delta) = 1$; this restriction ensures identifiability. Observe that, when both $\gamma$ and $\delta$ tend to infinity, the limiting distribution is multivariate normal; see more details in Barndorff-Nielsen [42].
Table 2. DIC values for the FMNIG-ME data.

| Model       | \( n = 100 \) | \( n = 500 \) |
|-------------|---------------|---------------|
| FMN-ME      | 970.2639      | 5209.1670     |
| FMT-ME      | 945.2123      | 4799.7490     |
| FMSL-ME     | 970.3589      | 4923.7360     |
| FMCN-ME     | 978.9807      | 4967.0020     |
| FSMN-ME     | 922.2715      | 4951.6620     |
| FMST-ME     | 874.2943      | 4546.9600     |
| FMSSL-ME    | 885.3027      | 4569.4180     |
| FMSCN-ME    | 882.9718      | 4578.2080     |

We now define an alternative ME model, by assuming that the marginal distribution of \( x \) is a finite mixture where the \( j \)th component is \( \text{NIG}(\mu_j, 1, \lambda_j, \gamma_j, \delta_j), j = 1, \ldots, G \). We call this the FMNIG-ME model, defined by

\[
Z | (x, U = u) \sim N_p (a + bx, u\Omega),
\]

\[
x | (U = u, S = j) \sim N(\mu_j + u\lambda_j, u),
\]

\[
U | S = j \sim IG(\gamma_j, \delta_j), \quad j = 1, \ldots, G,
\]

where \( \text{det}(\Omega) = 1 \). This definition is based on representation (13). The marginal distribution of \( Z \) is a mixture of NIG distributions, and each component \( j = 1, \ldots, G \) is distributed as \( \text{NIG}(a + b\mu_j, bb^\top + \Omega, (bb^\top + \Omega)^{-1}b\lambda_j, \gamma_j, \delta_j) \). Samples of size \( n = 100 \) and \( n = 500 \) from the FMNIG-ME model with \( p = 3 \) and \( G = 3 \) were generated with the following scenario: \( p_1 = 0.4, p_2 = 0.3, \mu_1 = -10, \mu_2 = 1, \mu_3 = 10, \lambda_1 = -2, \lambda_2 = 1, \lambda_3 = -2, \Omega = I_3, \alpha = (0.4, 0.1)^\top, \beta = (0.8, 0.9)^\top, \delta_1 = \delta_2 = \delta_3 = 0.5 \) and \( \gamma_1 = \gamma_2 = \gamma_3 = 1 \). Then we fitted FMSMSN-ME models with three components to these data. We considered models based on symmetric distributions, namely FMN-ME (normal), FMT-ME (Student-t), FMSL-ME (slash), FMCN-ME (contaminated normal) and the previously cited skewed models, namely, FSMN-ME, FMST-ME, FMSSL-ME and FMSCN-ME models. In each case, Table 2 presents the DIC values. It can be seen that the models that take into account skewness, heavy tails and multi-modality at the same time outperform (values in boldface) the other ones.

6.3. Simulation study 3 – identifiability

It is well known that identifiability is a sensitive issue for a mixture of regression models. Few works deal with this problem, and the results are restricted to some specific models. Extensions of these results to models based on the SMSM class are a challenge that, until now, has not been explored in depth in the literature. For a short discussion, see Zeller et al. [43]. Instead of a formal proof, we propose to study the identifiability of our proposed model using a simple method suggested by Lele et al. [44], called data cloning.

The data cloning algorithm allows us to approximate maximum likelihood estimates and the inverse of the Fisher information matrix using MCMC samples from a modified posterior distribution of the vector of parameters in the model, \( \Theta \). Let \( z = (z_1^\top, \ldots, z_n^\top)^\top \) be the observed sample and let \( z^{(K)} = (z_1^\top, \ldots, z_n^\top)^\top \) be the replicated data, which are obtained by replicating the original data \( K \) times.
Figure 3. Simulation study 3. Identifiability checking using data cloning.

The vector \( z^{(K)} \) is seen as a result of a hypothetical experiment that replicates the original one \( K \) times independently, yielding the same data \( z \) each time. Under suitable conditions, it is possible to show that, when \( K \) is large, the posterior distribution of \( \Theta|z^{(K)} \) is approximated by a normal distribution with mean equal to the maximum likelihood (ML) estimate \( \hat{\Theta} \) and covariance matrix equal to \( (1/K)I^{-1}(\hat{\Theta}) \), where \( I(\Theta) \) is the Fisher information matrix. Thus, the mean of MCMC samples drawn from the posterior distribution of \( \Theta|z^{(K)} \) can be used to approximate \( \hat{\Theta} \), and \( K \) times the covariance matrix of these posterior samples can be used to approximate the asymptotic covariance matrix of \( \hat{\Theta} \).

Also, Lele et al. [44] showed that if \( g(\Theta) \) is a function of the parameter vector \( \Theta \), and if the covariance matrix of the posterior distribution \( g(\Theta)|z^{(K)} \) has its largest eigenvalue \( \lambda_K \) converging to zero when \( K \) increases, then \( g(\Theta) \) is estimable. This convergence to zero has the same rate as \( 1/K \). Let \( \hat{\lambda}_K = \lambda_K/\lambda_1 \). The authors recommend detecting this convergence feature by the analysis of a plot of \( \hat{\lambda}_K \) as a function of \( K \), and comparing it with the expected value plot of \( 1/K \).

Figure 3 depicts these plots for the skewed models when \( g(\Theta) = \Theta \), where \( z \) is an artificial sample generated using the same setup of Section 6.1. The procedure was carried out using the R package dclone by Sólymos [45]. The plots suggest strong evidence of identifiability in all cases considered.

6.4. Simulation study – influence of atypical observations

The aim of this study is to analyse the influence of some atypical observations on the parameter estimates. It is similar to the study made by Massua et al. [46] in the context of censored linear regression models. We fixed \( r = 1 \) and \( G = 2 \) and generated one dataset of size \( n = 100 \) from the FMSN-ME model with the following parameter setup: \( \alpha = 0.4, \beta = 0.8, \omega_0^2 = 0.2, \omega_1^2 = 0.3, \mu_1 = 2, \mu_2 = 8, \Delta_1 = -2, \Delta_2 = 2, \gamma_1^2 = \gamma_2^2 = 0.1 \) and \( p_1 = 0.7 \). We perturbed three randomly chosen observations by increasing the values of the responses \( X \) and \( Y \) by \( \Lambda \% \) of their original value, for \( \Lambda = 10, 20, 30, \ldots, 150 \). For each of these 15 datasets, we fitted the FMSN-ME, FMST-ME and FMSSL-ME models and computed the relative change in the estimates of \( \alpha \) and \( \beta \) when compared with the fit of the original data. The relative change for \( \alpha \) is defined by \( \left| (\hat{\alpha} - \hat{\alpha}_I)/\hat{\alpha} \right| \times 100 \), where \( \hat{\alpha} \) and \( \hat{\alpha}_I \) denote posterior means before and after a set \( I \) of observations was perturbed, respectively. Analogously, we define the relative change for \( \beta \).
Table 3. Relative changes (in %) for $\alpha$ and $\beta$.

| Parameter | Model | FMSN-ME | FMST-ME | FMSSL-ME |
|-----------|-------|---------|---------|----------|
| $\alpha$  | $\Lambda$ |       |         |          |
| 50        |       | 5.7016  | 8.6901  | 7.7474   |
| 60        |       | 4.1320  | 9.5607  | 4.3337   |
| 70        |       | 2.4260  | 0.8962  | 1.5508   |
| 80        |       | 8.5715  | 0.3325  | 2.6586   |
| 90        |       | 7.5587  | 10.0411 | 0.8027   |
| 100       |       | 4.6819  | 0.8576  | 4.1425   |
| 110       |       | 21.0764 | 0.5531  | 9.0587   |
| 120       |       | 146.1072| 0.8381  | 0.7482   |
| 130       |       | 200.2993| 1.8805  | 0.6408   |
| 140       |       | 189.7507| 0.2774  | 1.4579   |
| 150       |       | 298.0920| 4.0190  | 5.1951   |
| $\beta$   |       | 0.7855  | 0.9880  | 0.1788   |
| 60        |       | 0.4573  | 0.3104  | 0.5077   |
| 70        |       | 0.2843  | 0.9337  | 0.4139   |
| 80        |       | 1.1458  | 0.1741  | 0.9156   |
| 90        |       | 0.4143  | 2.4703  | 0.0135   |
| 100       |       | 2.2378  | 1.0641  | 0.1198   |
| 110       |       | 1.8745  | 0.0238  | 1.5057   |
| 120       |       | 61.0258 | 0.0219  | 0.0462   |
| 130       |       | 170.4969| 0.3361  | 0.2159   |
| 140       |       | 162.9991| 1.0280  | 0.5672   |
| 150       |       | 171.6815| 0.9842  | 0.9270   |

The results are presented in Table 3 (only for $\Lambda \geq 50$). Regarding the parameter $\alpha$, we can see that for perturbations smaller than 110%, the relative changes are smaller than 11% and there is no precise pattern. But when $\Lambda$ becomes 110% or more, the relative changes for FMST-ME and FMSSL-ME models are under 10%, while for the FMSN-ME model it presents an increasing pattern, reaching almost 300% when $\Lambda = 150$. Regarding the parameter $\beta$, we can see that for $\Lambda$ smaller than 120% the three models behave in a very similar way, but, when $\Lambda$ increases, the FMSN-ME model loses performance, becoming less robust than the skewed models in dealing with outliers.

7. Analysis of a real dataset

We illustrate our proposed methods with the SLE data described in Section 1. In this case, $y$ and $x$ are the unobservable protein/creatinine ratio and 24-h proteinuria, respectively. The respective measurements taken from 75 patients are denoted by $Y$ and $X$. The main goal is to study the relationship between these two tests. Figure 1 shows that an FMSMSN-ME model with two components can be a proper choice to model these data. This is confirmed by the visual inspection of Figure 4.

We fitted FMSMSN-ME models with $G = 1$ and $G = 2$ to these data. In this example, we adopted a prior setup that was little different from that defined in Section 4, by fixing the hyperparameters of the prior distribution of $\rho$ in the FMCN-ME and FMSCN-ME models as $\rho_0 = \rho_1 = 2$. Considering the FMST-ME model with $G = 2$, running the Stan software on a 64-bit Windows environment on a laptop machine with a 1.80 GHz Intel Core i7 processor with 8.0 GB of RAM, the algorithm took 7.109709 min to obtain an MCMC sample of size 25,000.
Figure 4. Histograms for protein/creatinine ratio and 24-h proteinuria (both divided by 1000) for 75 SLE patients.

Table 4. Model selection for the SLE data.

| Model (G)     | DIC   | $\tau_D$ | log-lik  | LPML   | EAIC   | EBIC   |
|---------------|-------|----------|----------|--------|--------|--------|
| FMN-ME (1)    | 693.73 | 10.32    | -341.70  | 692.59 | 706.50 |        |
| FMST-ME (1)   | 494.63 | 6.29     | -244.17  | 496.68 | 515.22 |        |
| FMN-ME (2)    | 556.18 | 20.02    | -268.08  | 539.71 | 560.57 |        |
| FMT-ME (2)    | 544.71 | 33.23    | -255.62  | 509.02 | 532.19 |        |
| FMSL-ME (2)   | 553.17 | 19.78    | -266.69  | 540.39 | 563.56 |        |
| FMCN-ME (2)   | 561.00 | 26.44    | -267.28  | 534.75 | 560.24 |        |
| FMSN-ME (2)   | 490.17 | 12.09    | -239.04  | 488.42 | 513.91 |        |
| FMST-ME (2)   | 483.71 | 11.46    | -236.12  | 484.68 | 512.48 |        |
| FMSSL-ME (2)  | 489.37 | 11.24    | -239.06  | 490.26 | 518.07 |        |
| FMSNN-ME (2)  | 502.74 | 13.54    | -244.59  | 572.72 | 602.85 |        |

Table 4 presents the DIC, the effective dimension and the log-likelihood values (see Section 5) for the models. Also are presented the values of the LPML, EAIC and EBIC criteria. LPML is based on the conditional predictive ordinate (CPO) statistic for the $i$th observation, given by CPO$_i = \pi(z_i | z_{\neg i})$, that is, the posterior predictive density given all the observations excluding $z_i$. Then LPML = $\sum_{i=1}^{n} \log \text{CPO}_i$. EAIC and BIC are extensions of the usual AIC and BIC criteria to the Bayesian context [47]. These statistics can be easily approximated using posterior MCMC samples – for more details see Garay et al. [48]. According to all these criteria, the FMST-ME model with two components is the best one (values in boldface).

In order to study the fit of this model to the SLE data, we consider posterior predictive checking, by using the deviance $D(\cdot, \cdot)$ as a discrepancy measure between model and data – see Equation (18) – and computing the posterior predictive $p$-value (or Bayesian $p$-value), given by

$$p_B = P(D(w, \Theta) \geq D(z, \Theta) | z) = \int \int_A \pi(w | \Theta) \pi(\Theta | z) dw d\Theta,$$

where $z = (z_1^T, \ldots, z_n^T)^T$ is the observed sample, $w = (w_1^T, \ldots, w_n^T)^T$ is the replicated data that could have been observed and $A = \{(w, \Theta); \ D(w, \Theta) \geq D(z, \Theta)\}$. This $p$-value
is the posterior probability that a future observation is more extreme (as measured by the deviance) than the data; see Gelman et al. [49, sec. 6.3] for more details. Observe that \( p_B \) is computed with respect to the joint posterior distribution of \( (w, \Theta) \) given \( z \).

It is possible to approximate \( p_B \) using MCMC simulations. Let \( \Theta^{(l)} \) be the MCMC sample generated at the \( l \)th step of the algorithm, \( l = 1, \ldots, L \). Suppose that \( w^{(l)} \) is drawn from \( \pi(w \mid \Theta^{(l)}) \), which can be easily accomplished since this distribution is a mixture of SMSN distributions – see Equation (15). Then the pairs \( (w^{(l)}, \Theta^{(l)}), l = 1, \ldots, L \) are samples from the joint posterior distribution of \( w \) and \( \Theta \). Thus, to approximate the Bayesian \( p \)-value, it is enough to observe the relative frequency of the event \( A \) across the \( L \) samples, that is, the number of times \( D(w^{(l)}, \Theta^{(l)}) \) (the predictive deviance) exceeds \( D(z, \Theta^{(l)}) \) (the realized deviance) out of the \( L \) simulated draws. According to Gelman et al. [49], a model is suspect if a discrepancy is of practical importance and its \( p \)-value is close to 0 or 1. In the case of the FMST-ME model with two components, we obtained \( p_B \approx 0.5455 \), indicating no lack of fit at all. Additionally, Figure 5 shows a histogram of the differences \( D(w^{(l)}, \Theta^{(l)}) - D(z, \Theta^{(l)}) \) and a scatterplot of \( D(w^{(l)}, \Theta^{(l)}) \) by \( D(z, \Theta^{(l)}) \). Figure 6 presents a comparison between the actual data and some replicated data, showing a close agreement between them.

Table 5 shows the posterior estimates, standard deviations and 95% HPD intervals of the parameters in the FMST-ME model. In particular the credibility interval of \( \beta \) evidence a positive correlation between \( x \) and \( y \) and a difference in the relative calibration of the tests, which is the case when \( \beta \neq 1 \) [50]. Also, the use of a skewed and heavy-tailed model is justified by the estimates of \( \Delta_2 \) and \( \nu \), respectively. In the later case, observe that the posterior mean or mode are more suitable estimates, since the posterior distribution is highly skewed (see Figure 7).

According to Gelman et al. [51], in hierarchical models, when using the inverse-gamma(\( \xi, \xi \)) noninformative prior distribution for variance parameters, the resulting inferences can be sensitive to \( \xi \). In order to verify if our inferences are sensitive to the choice...
**Figure 6.** First row, first column: scatterplot of the SLE data. Other plots: scatterplots of replicated data.

**Figure 7.** SLE data. Histogram of the MCMC samples from the posterior distribution of $\nu$. 
Table 5. SLE data. Posterior estimates, standard deviations and 95% HPD intervals of the parameters in the FMST-ME model.

| Parameter | Mean  | Median | SD     | Lower  | Upper  |
|-----------|-------|--------|--------|--------|--------|
| $\beta$   | 0.6695| 0.6664 | 0.0404 | 0.5922 | 0.7492 |
| $\alpha$  | $-0.079$ | $-0.0732$ | 0.1617 | $-0.4016$ | 0.2211 |
| $\omega^2_1$ | 0.0108 | 0.0068 | 0.0145 | 0.0010 | 0.0291 |
| $\nu$     | 3.2754 | 2.2323 | 0.3964 | 1.5374 | 3.0552 |
| $\mu_1$   | 3.3268 | 3.3193 | 1.8707 | $-0.3718$ | 6.7285 |
| $\mu^2_2$ | 0.1364 | 0.1385 | 0.2567 | $-0.3426$ | 0.6548 |
| $\gamma^2_1$ | 1.8626 | 0.6618 | 2.6213 | 0.0065 | 7.3935 |
| $\gamma^2_2$ | 0.0156 | 0.0100 | 0.0191 | 0.0012 | 0.0423 |
| $\Delta_1$ | 3.0046 | 3.2066 | 2.0728 | $-1.8048$ | 6.8176 |
| $\Delta_2$ | 0.6879 | 0.6936 | 0.1996 | 0.2749 | 1.0701 |
| $r$       | 0.3044 | 0.2966 | 0.0802 | 0.1491 | 0.4618 |
| $\nu$     | 9.5523 | 5.8786 | 11.0247 | 0.9678 | 31.7304 |

of some hyperparameter values, we conducted, in the case of the real dataset analysis, a sensitivity analysis to evaluate the effect of prior choices on the estimates of the parameters. Recall we have adopted the following setup in the case of the dispersion parameters

$$\gamma_j^{-2} | f \sim \text{Gamma}(e,f) \quad j = 1, 2; \quad f \sim \text{Gamma}(g,h).$$

and $\omega_i^{-2} \sim \text{Gamma}(l,m), i = 0, 1$, where $e = 2$ and $g, h, l$ and $m$ have a common value $\xi$. Table A1 in the appendix shows the results for $\xi \in \{0.001, 0.1\}$. Although some discrepancies were observed in the estimates, the results, in general, resemble the original analysis (that is, with $\xi = 0.01$, Table 5) and did not change any conclusions regarding our best-fitted model.

8. Conclusion

In this article, we proposed an extension of the classical normal measurement error-in-variables model, flexible enough to accommodate at the same time skewness, heavy tails and multi-modality. Our approach is based on the joint modelling of the latent unobserved covariate and the random observational errors by a finite mixture of scale mixtures of skew-normal distributions. Stochastic representations of the model allow us to develop MCMC algorithms to perform Bayesian estimation of the parameters in the proposed model. Though the inspection of model selection criterion, simulated and real datasets were used to illustrate the advantages of our model over models based on symmetry. This method can be easily implemented using available software, making it useful for practitioners and researchers in several areas.

Computer Programs and Data

Computer programs and data are available at https://github.com/nelfilho?tab=repositories

Acknowledgments

The authors would like to thank an anonymous referee for insightful comments which substantially improved the paper. The research was supported by Universidade Federal do Amazonas (UFAM),
Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) and CNPq grants from the Brazilian federal government, and by FAPEAM grants from the government of the State of Amazonas, Brazil.

**Disclosure statement**

No potential conflict of interest was reported by the authors.

**ORCID**

Celso Rômulo Barbosa Cabral [http://orcid.org/0000-0001-6776-6690](http://orcid.org/0000-0001-6776-6690)

Nelson Lima de Souza [http://orcid.org/0000-0003-2325-6626](http://orcid.org/0000-0003-2325-6626)

Jeremias Leão [http://orcid.org/0000-0003-1176-0198](http://orcid.org/0000-0003-1176-0198)

**References**

[1] Fuller WA. Measurement error models. New York: John Wiley & Sons; 1987.

[2] Cheng CL, Van Ness JW. Statistical regression with measurement error. New York: John Wiley & Sons; 1999.

[3] Carroll RJ, Ruppert D, Stefanski LA, et al. Measurement error in nonlinear models: a modern perspective. Boca Raton: CRC Press; 2006.

[4] Rees F, Doherty M, Grainge MJ, et al. The worldwide incidence and prevalence of systemic lupus erythematosus: a systematic review of epidemiological studies. Rheumatology. 2017;56:1945–1961.

[5] Nieves CEF, Izmirly PM. Mortality in systemic lupus erythematosus: an updated review. Curr Rheumatol Rep. 2016;18:21.

[6] Lima DS. Estudo da relação proteína/creatinina em amostra isolada de urina x proteinúria de 24 horas na avaliação de pacientes com lúpus eritematoso sistêmico. Technical Report. Federal University of Amazonas; 2015.

[7] Bolfarine H, Arellano-Valle RB. Robust modelling in measurement error models using the t distribution. Brazilian J Probab Stat. 1994;8:67–84.

[8] Galea M, Bolfarine H, Vilca F. Local influence in comparative calibration models under elliptical t-distributions. Biometrical J: J Math Methods Biosci. 2005;47:691–706.

[9] de Castro M, Galea M. Robust inference in an heteroscedastic measurement error model. J Korean Stat Soc. 2010;39:439–447.

[10] Rocha GH, Loschi RH, Arellano-Valle RB. Bayesian mismeasurement t-models for censored responses. Statistics. 2016;50:841–869.

[11] Matos LA, Castro LM, Cabral CR, et al. Multivariate measurement error models based on Student-t distribution under censored responses. Statistics. 2018;52:1395–1416.

[12] Lachos V, Garibay V, Labra F, et al. A robust multivariate measurement error model with skew-normal/independent distributions and Bayesian MCMC implementation. Stat Methodol. 2009;6:527–541.

[13] Lachos V, Labra F, Bolfarine H, et al. Multivariate measurement error models based on scale mixtures of the skew–normal distribution. Statistics. 2010;44:541–556.

[14] Tomaya LC, de Castro M. A heteroscedastic measurement error model based on skew and heavy-tailed distributions with known error variances. J Stat Comput Simul. 2018;88:2185–2200.

[15] Carroll RJ, Roeder K, Wasserman L. Flexible parametric measurement error models. Biometrics. 1999;55:44–54.

[16] McLachlan G, Peel D. Finite mixture models. New York: John Wiley & Sons; 2000.

[17] Cabral CRB, Lachos VH, Prates MO. Multivariate mixture modeling using skew-normal independent distributions. Comput Stat Data Anal. 2012b;56:126–142.

[18] Cabral CRB, Lachos VH, Zeller CB. Multivariate measurement error models using finite mixtures of skew-student t distributions. J Multivar Anal. 2014;124:179–198.
[19] Birnbaum ZW. Effect of linear truncation on a multinormal population. Ann Math Stat. 1950;21:272–279.

[20] Nelson LS. The sum of values from a normal and a truncated normal distribution. Technometrics. 1964;6:469–471.

[21] Azzalini A. A class of distributions which includes the normal ones. Scand J Stat. 1985;12:171–178.

[22] Azzalini A, Dalla Valle A. The multivariate skew-normal distribution. Biometrika. 1996;83:715–726.

[23] Arellano-Valle RB, Azzalini A. On the unification of families of skew-normal distributions. Scand J Stat. 2006;33:561–574.

[24] Azzalini A, Capitanio A. The skew-normal and related families. Cambridge: Cambridge University Press; 2014.

[25] Arellano-Valle RB, Genton MG. On fundamental skew distributions. J Multivar Anal. 2005;96:93–116.

[26] Branco MD, Dey DK. A general class of multivariate skew-elliptical distributions. J Multivar Anal. 2001;79:99–113.

[27] Andrews DF, Mallows CL. Scale mixtures of normal distributions. J R Stat Soc, Ser B. 1974;36:99–102.

[28] R Core Team, R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2020.

[29] Dâvila VHL, Cabral CRB, Zeller CB. Finite mixture of skewed distributions. Cham: Springer; 2018.

[30] Vidal I, Castro LM. Influential observations in the independent student-t measurement error model with weak nondifferential error. Chil J Stat. 2010;1:17–34.

[31] Cabral CRB, Lachos VH, Madruga MR. Bayesian analysis of skew-normal independent linear mixed models with heterogeneity in the random-effects population. J Stat Plan Inference. 2012a;142:181–200.

[32] Richardson S, Green PJ. On Bayesian analysis of mixtures with an unknown number of components. J R Stat Soc, Ser B. 1997;59:731–792.

[33] Fonseca TCO, Ferreira MAR, Migon HS. Objective Bayesian analysis for the Student-t regression model. Biometrika. 2008;95:325–333.

[34] Garay AM, Bolfarine H, Lachos VH, et al. Bayesian analysis of censored linear regression models with scale mixtures of normal distributions. J Appl Stat. 2015;42:2694–2714.

[35] Congdon P. Bayesian statistical modelling. New York: John Wiley & Sons; 2007.

[36] Plummer M, et al. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In: Proceedings of the 3rd International Workshop on Distributed Statistical Computing. Vol. 124. Vienna, Austria; 2003, p. 1–10.

[37] Carpenter B, Gelman A, Hoffman MD, et al. Stan: a probabilistic programming language. J Stat Softw. 2017;76:1–32.

[38] Spiegelhalter DJ, Best NG, Carlin BP, et al. Bayesian measures of model complexity and fit. J R Stat Soc, Ser B. 2002;64:583–639.

[39] Stephens M. Bayesian methods for mixtures of normal distributions [Ph.D. thesis]. Oxford: Magdalen College; 1997.

[40] Celeux G, Forbes F, Robert CP, et al. Deviance information criteria for missing data models. Bayesian Anal. 2006;1:651–674.

[41] Spiegelhalter DJ, Best NG, Carlin BP, et al. The deviance information criterion: 12 years on. J R Stat Soc: Ser B: Stat Method. 2014;76:485–493.

[42] Barndorff-Nielsen OE. Normal inverse Gaussian distributions and stochastic volatility modelling. Scand J Stat. 1997;24:1–13.

[43] Zeller CB, Cabral CRB, Lachos VH, et al. Finite mixture of regression models for censored data based on scale mixtures of normal distributions. Adv Data Anal Classif. 2019;13:89–116.

[44] Lele SR, Nadeem K, Schmuland B. Estimability and likelihood inference for generalized linear mixed models using data cloning. J Am Stat Assoc. 2010;105:1617–1625.

[45] Sólymos P. dclone: data cloning in R. R Journal. 2010;2:29–37.
Appendices

Appendix 1. Sensitivity analysis from SLE data analysis

Table A1. Posterior point estimates and 95% confidence intervals for the parameters in the best-fitting FMST-ME model for the SLE data.

| Parameter | $\xi = 0.001$ | $\xi = 0.01$ |
|-----------|---------------|---------------|
| $\beta$   | Mean | Median | SD  | Lower | Upper | Mean | Median | SD  | Lower | Upper |
| $\alpha$  | 0.6720 | 0.6747 | 0.0331 | 0.6039 | 0.7418 | 0.6694 | 0.6670 | 0.0467 | 0.5897 | 0.7774 |
| $\omega_1^2$ | 0.0020 | 0.0011 | 0.0029 | 0.0000 | 0.0060 | 0.0010 | 0.0031 | 0.0361 | 0.0195 | 0.0097 | 0.1359 |
| $\omega_0^2$ | 2.2710 | 2.1715 | 0.4034 | 1.6976 | 3.1745 | 2.1508 | 2.1445 | 0.4795 | 1.4468 | 3.1874 |
| $\mu_1$   | 2.8326 | 2.8305 | 0.2699 | 2.3874 | 3.4497 | 2.4891 | 2.7247 | 0.9483 | 0.5480 | 3.9628 |
| $\mu_2^2$ | 0.9023 | 0.1299 | 0.2073 | 0.2674 | 0.4798 | 0.1909 | 0.1715 | 0.2590 | 0.2785 | 0.7388 |
| $\gamma_1^2$ | 0.0044 | 0.0018 | 0.0081 | 0.0000 | 0.0178 | 0.0250 | 0.0052 | 0.0682 | 0.0000 | 0.1168 |
| $\gamma_2^2$ | 0.0031 | 0.0020 | 0.0036 | 0.0000 | 0.0083 | 0.0114 | 0.0041 | 0.0188 | 0.0000 | 0.0500 |
| $\Delta_1$ | 3.8789 | 3.6619 | 0.8332 | 2.4202 | 5.4047 | 2.1423 | 4.1314 | 1.2169 | 2.1867 | 6.9904 |
| $\Delta_2$ | 0.7442 | 0.7549 | 0.1219 | 0.5255 | 0.9819 | 0.6345 | 0.6511 | 0.2409 | 0.1756 | 1.1004 |
| $\rho$    | 0.2739 | 0.2802 | 0.0622 | 0.1658 | 0.3698 | 0.2834 | 0.2826 | 0.0586 | 0.1718 | 0.3877 |
| $\nu$     | 9.3324 | 5.9511 | 9.7643 | 1.7127 | 28.8768 | 8.6442 | 5.6728 | 9.5183 | 1.2032 | 26.8021 |

The value $\xi$ denotes the hyperparameter choice corresponding to the inverse-gamma($\xi, \xi$) for the dispersion parameters.