JIRSS (2026)
Vol. 25, No. 2, pp 30-52
DOI: 10.22034/jirss.2026.2085508.1184

Maximum likelihood estimation in shape and skew scale mixtures of normal linear regression models: the impact of reparameterization

Zakaria Alizadeh Ghajari[1], Karim Zare[1], Soheil Shokri[2], Abdul Rasoul Ziaei[3]

[1]Department of Statistics, Marv.C., Islamic Azad University, Marvdasht, Iran.
[2]Department of Mathematics and Statistics, Ra.C., Islamic Azad University, Rasht, Iran.
[3]Department of Statistics, Das.C., Islamic Azad University, Borazjan, Iran.

Received: 2026-02-15, Accepted: 2026-07-25, Published online: 2026-07-30

Abstract. This study introduces a novel class of Linear Regression Models (LRM), the Shape and Skew Scale Mixtures of Normal (SHSSMN) distributions, to effectively handle skewed and heavy-tailed error terms. By extending the Expectation-Maximization (EM) algorithm, we explore two distinct scenarios based on without or with reparametrization. Simulation results demonstrate that the parameter estimates under both scenarios are consistent and nearly identical, with the second scenario offering a computational advantage by reducing the number of iterations required for convergence, especially for complex distributions. The models are applied to real-world data from the Australian Institute of Sport (AIS) to evaluate their practical utility. The results show that the SHSSMN models, particularly the Modified Skew-t Cauchy (STC) distribution, provide superior fit and effectively address the skewness and heavy-tailed nature of the residuals, making them a valuable tool for regression analysis in diverse fields.

Keywords. EM-type algorithm, Linear regression models, Shape and skew scale mixtures of normal distributions.

MSC: 62F10, 62J05, 65C05.

Zakaria Alizadeh Ghajari (zakariaalizadeh@iau.ac.ir).
Corresponding Author: Karim Zare (karim.zare@iau.ac.ir).
Soheil Shokri (soheilshokri@iau.ac.ir).
Abdul Rasoul Ziaei (ar.ziaei@iau.ac.ir).

1 Introduction

An effective approach to enhancing the flexibility of statistical modeling for skewed and heavy-tailed data is the use of the skew family of distributions. This family includes distributions characterized by skewness and heavy tails (Arellano-Valle et al., 2018; Arnold and Beaver, 2000; Azzalini and Capitanio, 2003, 1999; Branco and Dey, 2001; Genton and Loperfido, 2005; Gómez et al., 2007; Gupta et al., 2002; Nadarajah and Kotz, 2003; Nematollahi et al., 2016; Vilca et al., 2014; Wang et al., 2004; Wang and Genton, 2006). Within this category, the Skew-Normal (SN) family [refer to Section 2] represents a prominent subset with distinctive properties. In recent years, researchers have widely applied the SN family in various statistical frameworks, particularly in regression models (Arellano-Valle et al., 2004; Ferreira et al., 2015; Ferreira, 2008; Alizadeh Ghajari et al., 2024; Mattos et al., 2018; Sahu et al., 2003). The fundamental component of this family is the SN distribution, derived from the Normal distribution. To clarify the methodology used in this study, we provide a formal definition of the SN distribution, presented by Azzalini (1985).

Let Y be a random variable that follows a SN distribution, characterized by a location parameter $\mu \in \mathbb{R}$, a scale parameter $\sigma^2 > 0$, and a shape (or skewness) parameter $\lambda \in \mathbb{R}$. This is compactly represented as $Y \sim SN(\mu, \sigma^2, \lambda)$. The Probability Density Function (PDF) of SN distribution for all $y \in \mathbb{R}$ is expressed as:

$$f(y; \mu, \sigma^2, \lambda) = 2\phi(y; \mu, \sigma^2)\Phi\left(\lambda \frac{y-\mu}{\sigma}\right),$$

where $\phi(.; \mu, \sigma^2)$ denotes the PDF of a normal distribution with mean μ and variance σ^2, and $\Phi(.)$ is the cumulative distribution function (CDF) of the standard normal distribution.

Estimating the parameters of the SN distribution using the maximum likelihood (ML) method presents significant computational challenges. These arise primarily from the complexity of the log-likelihood function, due to the presence of the term $\Phi(.)$, which involves the integral of $\phi(.)$. Although approximations for $\Phi(.)$ have been proposed by Ashour and Abdel-hameed (2010); Watagoda et al. (2019), their discrete nature further complicates the derivation process. To address this issue, a latent variable U is introduced into the joint density function $f(y, u)$. This incorporation facilitates the use of the complete log-likelihood function, which depends on both observed and unobserved (latent) data, along with the parameters of the model. The next step involves calculating the conditional expectation of the complete log-likelihood function, given the observed data. This leads to the construction of the Q-function, which is then optimized with respect to the parameters. This iterative process forms the basis of the Expectation-Maximization (EM) algorithm. As highlighted by Dempster et al. (1977), the EM algorithm has become a cornerstone for parameter estimation in models with incomplete data. For further details, refer to Celeux (1985); Delyon et al. (1999); Kuhn and Lavielle (2004); Meng and Rubin (1993); Wei and Tanner (1990).

To derive the density function of the observed and latent variables (the compete variables) for the SN distribution, two primary scenarios have been explored in prior

research.

First scenario:
In this scenario, the joint density function of Y and the latent variable W is defined as:

$$f_1(y,w) = 2\phi(y;\mu,\sigma^2)\phi(w;\lambda(y-\mu),\sigma^2),$$

a formulation frequently cited in the literature (Arellano-Valle et al., 2009; Ferreira et al., 2011, 2015; Alizadeh Ghajari et al., 2024; Kahrari et al., 2016). This relationship can be derived through two distinct methods:

Method 1: Hierarchical Representation
The hierarchical representation provides a structured framework for obtaining the joint density function. Specifically, the SN distribution can be expressed hierarchically as follows:
If $Y \sim \text{SN}(\mu,\sigma^2,\lambda)$, then:

$$Y|u \sim \text{N}\left(\mu + \frac{\sigma\lambda u}{\sqrt{1+\lambda^2}}, \frac{\sigma^2}{\sqrt{1+\lambda^2}}\right),$$

where $U \sim \text{HN}(0,1)$ and HN(.) represents the Half-Normal distribution. In this framework, the conditional distribution of Y given $U = u$ is normal, and $f_1(y,w)$ is obtained through the change of variable $W = \sqrt{1+\lambda^2}\sigma U$.

Method 2: Based on the Definition of $\Phi(.)$
Using the definition of $\Phi(.)$, we have:

$$\Phi\left(\lambda\frac{y-\mu}{\sigma}\right) = \int_{-\infty}^{\frac{\lambda(y-\mu)}{\sigma}} \phi(z;0,1)\text{d}z.$$

By applying the properties of the normal distribution and a change of variables, this can be rewritten as:

$$\Phi\left(\lambda\frac{y-\mu}{\sigma}\right) = \int_{0}^{+\infty} \phi(w;\lambda(y-\mu),\sigma^2)\text{d}w.$$

Consequently, the marginal density function of Y is given by:

$$f_1(y) = 2\phi(y;\mu,\sigma^2)\int_{0}^{+\infty} \phi(w;\lambda(y-\mu),\sigma^2)\text{d}w,$$

leading to the desired result, $f_1(y,w)$.

Second scenario:
In this scenario, the joint density function of Y and W is defined as:

$$f_2(y,w) = 2\phi(y;\mu,\sigma^2)\phi\left(w;\lambda\frac{y-\mu}{\sigma},1\right).$$

By deriving the corresponding Q with respect to σ^2 or σ results in a quadratic equation in σ, which lacks a closed-form solution. To address this challenge, reparameterization is introduced by defining Δ as, $\Delta = \frac{\lambda}{\sigma}$, we have

$$f_2(y, w) = 2\phi(y; \mu, \sigma^2)\phi(w; \Delta(y - \mu), 1).$$

a formulation discussed in Arellano-Valle et al. (2018). As in the first scenario, the marginal density function of Y can be derived either from the hierarchical representation, using the change of variable $W = \sqrt{1 + \lambda^2}U$, or from the definition of $\Phi(.)$; both methods lead to $f_2(y, w)$.

To the best of our knowledge, no previous study has systematically compared the effects of these two scenarios on EM-based estimation across the broader family of SN distributions and related models, particularly for complex structures such as SHSSMN-LRM. This reparameterization can simplify the M-step, improve computational stability, and reduce the number of iterations required for convergence, especially in complex models. It may also offer theoretical advantages, including easier computation of the observed information matrix and a more convenient framework for diagnostic analysis.

The present study has two main objectives: (i) to develop EM-based procedures for parameter estimation and computation of the information matrix for SHSSMN-LRM under both scenarios, and (ii) to compare the two scenarios in terms of their theoretical properties and computational performance. To address these objectives, the remainder of the paper is structured as follows: Section 2 introduces the rich and flexible SHSSMN family as defined by Arellano-Valle et al. (2018). The mixture structure of this distribution is defined through two latent variables, which control the scale and shape components of the conditional SN distribution. Section 3 details the SHSSMN-LRM and the application of the EM algorithm under two specific scenarios. Also, the observed information matrix is employed to compute the asymptotic covariance of the ML estimates. Section 4 presents simulation results and a real-world data application. Finally, Section 5 concludes with key insights from the study and outlines potential directions for future research.

2 SHSSMN distribution

Here, we adopt the definition of the SHSSMN distribution as outlined in Arellano-Valle et al. (2018). A random variable Y follows a SHSSMN distribution characterized by a location parameter $\mu \in \mathbb{R}$, a scale parameter $\sigma^2 > 0$, and a shape parameter $\lambda \in \mathbb{R}$ if there exist two random variables V and S, with the joint density $q(v, s|\nu, \tau)$ such that

$$Y|v, s \sim \text{SN}\left(\mu, \sigma^2\kappa(v), \lambda\rho(s)\sqrt{\kappa(v)}\right),$$

for some scale function $\kappa(v)$ and real-valued shape function represented by $\rho(s)\sqrt{\kappa(v)}$ where the distribution of shape function is not symmetric about zero. This distribution is briefly denoted by $Y \sim \text{SHSSMN}(\mu, \sigma^2, \lambda, \nu, \tau)$. In this representation, $\kappa(v)$ determines

the scale-mixture component, while $\rho(s)\sqrt{\kappa(v)}$ determines the shape-mixture component. Hence, the term "shape mixtures" refers to the mixing induced by $\rho(s)\sqrt{\kappa(v)}$, and "scale mixtures" refers to the mixing induced by $\kappa(v)$.
The conditional PDF of Y given v and s is obtained as follows:

$$f(y|v,s) = 2\phi(y;\mu,\sigma^2\kappa(v))\Phi\left(\lambda\rho(s)\sqrt{\kappa(v)}\frac{y-\mu}{\sqrt{\sigma^2\kappa(v)}}\right)$$
$$= 2\phi(y;\mu,\sigma^2\kappa(v))\Phi\left(\lambda\rho(s)\frac{y-\mu}{\sigma}\right).$$

Based on the above relationship, the joint density of Y, V and S is:

$$f(y,v,s) = 2\phi(y;\mu,\sigma^2\kappa(v))\Phi\left(\lambda\rho(s)\frac{y-\mu}{\sigma}\right)q(v,s). \tag{2.1}$$

Then, the PDF of Y is:

$$f(y) = 2\int\int \phi(y;\mu,\sigma^2\kappa(v))\Phi\left(\lambda\rho(s)\frac{y-\mu}{\sigma}\right)q(v,s)\mathrm{d}v\mathrm{d}s.$$

Assuming that V and S are independent random variables, the joint density satisfies $q(v,s) = h(v)g(s)$ and hence the PDF of Y can be expressed as

$$f(y) = 2\int \phi\left(y;\mu,\sigma^2\kappa(v)\right)h(v)\mathrm{d}v\int \Phi\left(\lambda\rho(s)\frac{y-\mu}{\sigma}\right)g(s)\mathrm{d}s.$$

where $h(v)$ and $g(s)$ denote the PDFs of V and S, respectively.
The models listed below are obtained as special cases by selecting specific forms of $\kappa(v)$, $\rho(s)$, and the joint distribution of V and S.

1. if $\kappa(v) = \rho(s) = 1$ then $Y \sim \mathrm{SN}(\mu,\sigma^2,\lambda)$.
2. if $\rho(s) = 1$ then $Y|\ v \sim \mathrm{SN}(\mu,\sigma^2\kappa(v),\lambda\sqrt{\kappa(v)})$ and we have skew scale mixtures of normal (SSMN) distribution (Ferreira et al., 2011).
    a. if $\kappa(v) = \frac{1}{v}$ and $V \sim \mathrm{Gamma}(\frac{\nu}{2},\frac{\nu}{2})$ we have skew-t- normal (STN) distribution with the PDF as follows:
    $$f(y) = 2t(y;\mu,\sigma^2,\nu)\Phi\Big(\lambda\frac{y-\mu}{\sigma}\Big).$$
    b. if $\kappa(v) = \frac{1}{v}$ and $V \sim \mathrm{Beta}(\nu,1)$, we have skew-slash-normal (SSLN) distribution. The pdf of Y is:
    $$f(y) = 2sl(y;\mu,\sigma^2,\nu)\Phi\Big(\lambda\frac{y-\mu}{\sigma}\Big).$$
3. if $\kappa(v) = 1$ and $Y|s \sim \mathrm{SN}(\mu,\sigma^2,\lambda\rho(s))$, we have shape mixtures of skew-normal (SHMSN) distribution.

a. if $\rho(s) = s$ and $S \sim \mathrm{N}(1, \tau)$, we have skew-generalized- normal (SGN) distribution (Arellano-Valle et al., 2004). The PDF of Y is:

$$f(y) = 2\phi(y; \mu, \sigma^2)\Phi\left(\frac{\lambda\frac{y-\mu}{\sigma}}{\sqrt{1+\tau(\lambda\frac{y-\mu}{\sigma})^2}}\right).$$

b. if $\rho(s) = s$ and $S \sim \mathrm{HN}(0, 1)$, we have skew-normal- Cauchy (SNC) distribution (Kahrari et al., 2016). The PDF of Y is:

$$f(y) = 2\phi(y; \mu, \sigma^2)C(\lambda\frac{y-\mu}{\sigma}),$$

where $C(.)$ is the the CDF of the standard Cauchy distribution.

4. if $\kappa(v) = \frac{1}{v}$, $\rho(s) = s$ and V, S are independent:

a. if $V \sim \mathrm{Gamma}(\frac{\nu}{2}, \frac{\nu}{2})$ and $S \sim \mathrm{N}(1, 1)$, we have Modified skew-t-normal (MSTN) distribution. The PDF of Y is:

$$f(y) = 2t(y; \mu, \sigma^2, \nu)\Phi\left(\frac{\lambda\frac{y-\mu}{\sigma}}{\sqrt{1+(\lambda\frac{y-\mu}{\sigma})^2}}\right).$$

b. if $V \sim \mathrm{Beta}(\nu, 1)$ and $S \sim \mathrm{N}(1, 1)$, we have Modified skew-slash-normal (MSSLN) distribution. The PDF of Y is:

$$f(y) = 2sl(y; \mu, \sigma^2, \nu)\Phi\left(\frac{\lambda\frac{y-\mu}{\sigma}}{\sqrt{1+(\lambda\frac{y-\mu}{\sigma})^2}}\right),$$

where $sl(.; \mu, \sigma^2, \nu)$ is the PDF of of the Slash distribution.

c. if $V \sim \mathrm{Gamma}(\frac{\nu}{2}, \frac{\nu}{2})$ and $S \sim \mathrm{HN}(0, 1)$, we have STC distribution. The PDF of Y is:

$$f(y) = 2t(y; \mu, \sigma^2, \nu)C(\lambda\frac{y-\mu}{\sigma}).$$

d. if $V \sim \mathrm{Beta}(\nu, 1)$ and $S \sim \mathrm{HN}(0, 1)$, we have Modified skew-slash-Cauchy (SSLC) distribution. The PDF of Y is:

$$f(y) = 2sl(y; \mu, \sigma^2, \nu)C(\lambda\frac{y-\mu}{\sigma}).$$

3 Model and maximum likelihood estimation

3.1 SHSSMN linear regression models

Consider the following SHSSMN-LRM model:

$$Y_i = \boldsymbol{x}_i^\top \beta + \varepsilon_i, \qquad i = 1, 2, \ldots, n,$$

where $\varepsilon_i \sim \text{SHSSMN}(0, \sigma^2, \lambda, \nu, \tau)$. In this equation, Y_i represents the i-th response variable, $\boldsymbol{x}_i$ denotes a known $p \times 1$ explanatory vector, β signifies a p-dimensional vector of unknown regression coefficients, and ε_i denotes the i-th random error term. It can be obviously shown that if $\varepsilon_i \sim \text{SHSSMN}(0, \sigma^2, \lambda, \nu, \tau)$ then:

$$Y_i \sim \text{SHSSMN}(\boldsymbol{x}_i^\top \beta, \sigma^2, \lambda, \nu, \tau).$$

From equation (2.1), the joint density of Y_i, V_i and S_i is expressed as:

$$f(y_i, v_i, s_i) = 2\phi(y_i; \boldsymbol{x}_i^\top \beta, \sigma^2 \kappa(v_i))\Phi\left(\lambda \rho(s_i) \frac{y_i - \boldsymbol{x}_i^\top \beta}{\sigma}\right) h(v_i) g(s_i). \tag{3.1}$$

Now, we examine the subsequent two scenarios to estimate the parameters utilizing the EM algorithm.
As before, in the first scenario, the following integral representation is used for Φ:

$$\Phi\left(\lambda \rho(s_i) \frac{y_i - \boldsymbol{x}_i^\top \beta}{\sigma}\right) = \int_0^{+\infty} \phi(w_i; \lambda \rho(s_i)(y_i - \boldsymbol{x}_i^\top \beta), \sigma^2) \mathrm{d}w_i.$$

Substituting this into equation (3.1), the joint density of the random variables Y_i, V_i, S_i and W_i becomes:

$$f_1(y_i, v_i, s_i, w_i) = 2\phi(y_i; \boldsymbol{x}_i^\top \beta, \sigma^2 \kappa(v_i))\phi(w_i; \lambda \rho(s_i)(y_i - \boldsymbol{x}_i^\top \beta), \sigma^2) h(v_i) g(s_i).$$

This representation introduces the latent variable w, which can be incorporated into the EM algorithm to enhance the parameter estimation process under this scenario.
Subsequently, we define the complete data for the $i-th$ individual as $d_i = (y_i, v_i, s_i, w_i)^\top$, where y_i denotes the observed data point and (v_i, s_i, w_i) are the corresponding unobserved latent variables. The complete data for n individuals is then represented as $\boldsymbol{d} = (d_1, \ldots, d_n)^\top$.
The log-likelihood function for the parameter vector $\theta = (\beta^\top, \sigma^2, \lambda, \nu, \tau)^\top$, based on the complete data $\boldsymbol{d}$, is denoted by $l_{1c}(\theta|\boldsymbol{d})$ and is given by

$$l_{1c}(\theta|\boldsymbol{d}) = \sum_{i=1}^{n} \log f_1(y_i, v_i, s_i, w_i),$$

which expands to:

$$\begin{aligned} l_{1c}(\theta|\boldsymbol{d}) = \sum_{i=1}^{n} \ln \phi(y_i; \boldsymbol{x}_i^\top \beta, \sigma^2 \kappa(v_i)) + \sum_{i=1}^{n} \ln \phi(w_i; \lambda \rho(s_i)(y_i - \boldsymbol{x}_i^\top \beta), \sigma^2) \\ + \sum_{i=1}^{n} \ln h(v_i) + \sum_{i=1}^{n} \ln g(s_i) + n \ln 2. \end{aligned}$$

The final term does not contribute to the estimation of θ and can therefore be omitted during the implementation of the EM algorithm.
The E-step on the $(r+1)$-th iteration requires the calculation of the conditional expectation of the complete log-likelihood function given the observed data (Q-displacement function) outlined below as:

$$Q_1(\theta|\theta^{(r)}) \propto \sum_{i=1}^{n} Q_{1i}^{(r)},$$

where

$$\begin{aligned} Q_{1i}^{(r)} = &- \ln \sigma^2 - \frac{1}{2\sigma^2} h_{1i}^{(r)} (y_i - \boldsymbol{x}_i^\top \beta)^2 - \frac{1}{2\sigma^2} \widehat{w_{1i}^2}^{(r)} + \frac{\lambda}{\sigma^2} \widehat{w\rho_{1i}}^{(r)} (y_i - \boldsymbol{x}_i^\top \beta) \\ &+ E_1\{\ln h(V_i)|y_i, \theta^{(r)}\} + E_1\{\ln g(S_i)|y_i, \theta^{(r)}\}, \end{aligned}$$

in which,

$$\begin{aligned} h_{1i}^{(r)} &= \widehat{\kappa_{1i}^{-1}}^{(r)} + \lambda^{2(r)} \widehat{\rho_{1i}^2}^{(r)}, \\ \widehat{\kappa_{1i}^{-1}}^{(r)} &= E_1\left(\kappa^{-1}(V_i)|y_i, \theta^{(r)}\right), \\ \widehat{w_{1i}^2}^{(r)} &= E_1\left(W_i^2|y_i, \theta^{(r)}\right), \\ \widehat{w\rho_{1i}}^{(r)} &= E_1\left(W_i \rho(s_i)|y_i, \theta^{(r)}\right), \\ \widehat{\rho_{1i}^2}^{(r)} &= E_1\left(\rho^2(S_i)|y_i, \theta^{(r)}\right), \end{aligned}$$

and $\theta^{(r)} = \left(\beta^{(r)\top}, \sigma^{2(r)}, \lambda^{(r)}, \nu^{(r)}, \tau^{(r)}\right)^\top$ denotes the parameter estimates from the r-th iteration.

Next, the Q_1-function is maximized with respect to θ to obtain the updated parameter estimates, $\theta^{(r+1)}$. The M-step of the EM algorithm is carried out as follows:

(1) Update $\beta^{(r)}$ by

$$\begin{aligned} \beta^{(r+1)} = &\left(\boldsymbol{X}^\top \left(D\left(\widehat{\kappa_1^{-1}}^{(r)}\right) + \lambda^{2(r)} D\left(\widehat{\rho_1^2}^{(r)}\right)\right)\boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \\ &\left(D\left(\widehat{\kappa_1^{-1}}^{(r)}\right)\boldsymbol{y} - \lambda^{(r)} \widehat{w\rho_1}^{(r)} + \lambda^{2(r)} D\left(\widehat{\rho_1^2}^{(r)}\right)\boldsymbol{y}\right), \end{aligned}$$

where $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)^\top$ is the matrix of regressors and $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^\top$ is the response vector, and $\widehat{w\rho_1}^{(r)} = \left(\widehat{w\rho_{11}}^{(r)}, \ldots, \widehat{w\rho_{1n}}^{(r)}\right)^\top$. Moreover, $D(\boldsymbol{t}) = Diag(t_1, \ldots, t_n)$ for a vector $\boldsymbol{t} = (t_1, \ldots, t_n)^\top$.

(2) Update $\sigma^{2(r)}$ by

$$\begin{aligned}\sigma^{2(r+1)} =& \frac{1}{2n}\Big((\boldsymbol{y}-\boldsymbol{X}\beta^{(r+1)})^\top D\Big(\widehat{\kappa_1^{-1}}^{(r)}\Big)(\boldsymbol{y}-\boldsymbol{X}\beta^{(r+1)}) + 1_n^\top \widehat{w_1^2}^{(r)} \\ & - 2\lambda^{(r)}\widehat{w\rho_1}^{(r)\top}(\boldsymbol{y}-\boldsymbol{X}\beta^{(r+1)}) \\ & + \lambda^{2(r)}(\boldsymbol{y}-\boldsymbol{X}\beta^{(r+1)})^\top D\Big(\widehat{\rho_1^2}^{(r)}\Big)(\boldsymbol{y}-\boldsymbol{X}\beta^{(r+1)})\Big),\end{aligned}$$

where 1_n denotes a vector of ones with length n.

(3) Update $\lambda^{(r)}$ by

$$\lambda^{(r+1)} = \frac{\widehat{w\rho_1}^{(r)\top}(\boldsymbol{y}-\boldsymbol{X}\beta^{(r+1)})}{(\boldsymbol{y}-\boldsymbol{X}\beta^{(r+1)})^\top D\Big(\widehat{\rho_1^2}^{(r)}\Big)(\boldsymbol{y}-\boldsymbol{X}\beta^{(r+1)})}.$$

(4) Update $\nu^{(r)}$ by

$$\nu^{(r+1)} = \text{argmax}_\nu \left[\sum_{i=1}^n E_1(\ln h(V_i)|y_i)\right].$$

(5) Update $\tau^{(r)}$ by

$$\tau^{(r+1)} = \text{argmax}_\tau \left[\sum_{i=1}^n E_1(\ln g(S_i)|y_i)\right].$$

In situations where M-steps (4 and 5) are challenging to compute, the following ML-steps can be used as alternatives.

(1) Update $\nu^{(r)}$ by

$$\nu^{(r+1)} = \text{argmax}_\nu \left[\sum_{i=1}^n \ln \int \phi(y_i; \boldsymbol{x}_i^\top \beta^{(r+1)}, \sigma^{2(r+1)}\kappa(v_i))h(v_i)\mathrm{d}v_i\right].$$

(2) Update $\tau^{(r)}$ by

$$\tau^{(r+1)} = \text{argmax}_\tau \left[\sum_{i=1}^n \ln \int \Phi(\lambda^{(r+1)}\rho(s_i)\frac{y_i - \boldsymbol{x}_i^\top \beta^{(r+1)}}{\sigma^{(r+1)}})g(s_i)\mathrm{d}s_i\right].$$

The conditional expectations are computed as below:

$$E_1(\kappa^{-1}(V_i)|y_i) = \frac{\int \kappa^{-1}(v_i)\phi(y_i; \boldsymbol{x}_i^\top\beta, \sigma^2\kappa(v_i))h(v_i)\mathrm{d}v_i}{\int \phi(y_i; x_i^\top\beta, \sigma^2\kappa(v_i))h(v_i)\mathrm{d}v_i},$$

$$E_1(\rho^2(S_i)|y_i) = \frac{\int \rho^2(s_i)\Phi\left(\lambda\rho(s_i)\frac{y_i-\boldsymbol{x}_i^\top\beta}{\sigma}\right)g(s_i)\mathrm{d}s_i}{\int \Phi\left(\lambda\rho(s_i)\frac{y_i-\boldsymbol{x}_i^\top\beta}{\sigma}\right)g(s_i)\mathrm{d}s_i},$$

$$\begin{aligned} E_1(W_i^2|y_i) &= \lambda^2(y_i-\boldsymbol{x}_i^\top\beta)^2E_1(\rho^2(S_i)|y_i) + \sigma^2 \\ &\quad + \lambda(y-\boldsymbol{x}^\top\beta)\sigma\frac{\int \rho(s_i)\phi\left(\lambda\rho(s_i)\frac{y_i-\boldsymbol{x}_i^\top\beta}{\sigma}\right)g(s_i)\mathrm{d}s_i}{\int \Phi\left(\lambda\rho(s_i)\frac{y_i-\boldsymbol{x}_i^\top\beta}{\sigma}\right)g(s_i)\mathrm{d}s_i}, \\ E_1(W\rho(S_i)|y_i) &= E_1(\rho(S_i)E_1(W|y_i,s_i)|y_i) \\ &= E_1\left(\rho(S_i)\left(\lambda\rho(S_i)(y_i-\boldsymbol{x}_i^\top\beta) + \sigma\frac{\phi\left(\lambda\rho(S_i)\frac{y_i-\boldsymbol{x}_i^\top\beta}{\sigma}\right)}{\Phi\left(\lambda\rho(S_i)\frac{y_i-\boldsymbol{x}_i^\top\beta}{\sigma}\right)}\right)\middle| y_i\right). \end{aligned}$$

Since $W|y_i, s_i \sim TN(\lambda\rho(s_i)(y_i-\boldsymbol{x}_i^\top\beta), \sigma^2, (0,+\infty))$, this simplifies to:

$$E_1(W\rho(S_i)|y_i) = \lambda(y_i-\boldsymbol{x}_i^\top\beta)E_1(\rho^2(S_i)|y_i) + \sigma\frac{\int \rho(s_i)\phi\left(\lambda\rho(s_i)\frac{y_i-\boldsymbol{x}_i^\top\beta}{\sigma}\right)g(s_i)\mathrm{d}s_i}{\int \Phi\left(\lambda\rho(s_i)\frac{y_i-\boldsymbol{x}_i^\top\beta}{\sigma}\right)g(s_i)\mathrm{d}s_i}.$$

where, TN denotes the truncated normal distribution.
In the second scenario, we use the following equation:

$$\Phi\left(\lambda\rho(s_i)\frac{y_i-\boldsymbol{x}_i^\top\beta}{\sigma}\right) = \int_0^{+\infty}\phi(w_i; \Delta\rho(s_i)(y_i-\boldsymbol{x}_i^\top\beta), 1)\mathrm{d}w_i.$$

In this case, the joint density of the random variables Y_i, V_i, S_i and W_i is

$$f_2(y_i, v_i, s_i, w_i) = 2\phi(y_i; \boldsymbol{x}_i^\top\beta, \sigma^2\kappa(v_i))\phi(w_i; \Delta\rho(s_i)(y_i-\boldsymbol{x}_i^\top\beta), 1)h(v_i)g(s_i).$$

Therefore, the log-likelihood function for the parameter vector $\theta = (\beta^\top, \sigma^2, \Delta, \nu, \tau)^\top$, based on the complete data $\boldsymbol{d}$, is denoted by $l_{2c}(\theta|\boldsymbol{d})$ and is given by

$$l_{2c}(\theta|\boldsymbol{d}) = \sum_{i=1}^{n}\log f_2(y_i, v_i, s_i, w_i).$$

which expands is proportional to:

$$\begin{aligned} l_{2c}(\theta|\boldsymbol{d}) \propto &\sum_{i=1}^{n}\ln\phi(y_i; \boldsymbol{x}_i^\top\beta, \sigma^2\kappa(v_i)) + \sum_{i=1}^{n}\ln\phi(w_i; \Delta\rho(s_i)(y_i-\boldsymbol{x}_i^\top\beta), 1) \\ &+ \sum_{i=1}^{n}\ln h(v_i) + \sum_{i=1}^{n}\ln g(s_i), \end{aligned}$$

and hence, the Q_2-function is given by

$$Q_2(\theta|\theta^{(r)}) \propto \sum_{i=1}^{n} Q_{2i}^{(r)},$$

where

$$\begin{aligned} Q_{2i}^{(r)} = &-\frac{1}{2}\ln\sigma^2 - \frac{1}{2}h_{2i}^{(r)}(y_i - \boldsymbol{x}_i^\top\beta)^2 + \Delta\widehat{w\rho_{2l}}^{(r)}(y_i - \boldsymbol{x}_i^\top\beta) - \frac{1}{2}\sum_{i=1}^{n} E_2(W_i^2|y_i) \\ &+ E_2\{\ln h(V_i)|y_i, \theta^{(r)}\} + E_2\{\ln g(S_i)|y_i, \theta^{(r)}\}, \end{aligned}$$

in which,

$$\begin{aligned} h_{2i}^{(r)} &= \frac{1}{\sigma^2}\widehat{\kappa_{2i}^{-1}}^{(r)} + \Delta^{2(r)}\widehat{\rho_{2i}^2}^{(r)}, \\ \widehat{\kappa_{2i}^{-1}}^{(r)} &= E_2\left(\kappa^{-1}(V_i)|y_i, \theta^{(r)}\right), \\ \widehat{w\rho_{2i}}^{(r)} &= E_2\left(W_i\rho(S_i)|y_i, \theta^{(r)}\right), \end{aligned}$$

and

$$\widehat{\rho_{2i}^2}^{(r)} = E_2\left(\rho^2(S_i)|y_i, \theta^{(r)}\right).$$

When differentiating the Q_2 function, the term $\sum_{i=1}^{n} E_2(W_i^2|y_i)$ can be disregarded, as it remains independent of the unknown parameters.
Based on the above, The M-steps of EM algorithm are as follows:

(1) Update $\beta^{(r)}$ by

$$\begin{aligned} \beta^{(r+1)} = &\left(\boldsymbol{X}^\top\left(\frac{1}{\sigma^{2(r)}}D\left(\widehat{\kappa_2^{-1}}^{(r)}\right) + \Delta^{2(r)}D\left(\widehat{\rho_2^2}^{(r)}\right)\right)\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top \\ &\left(\frac{1}{\sigma^{2(r)}}D\left(\widehat{\kappa_2^{-1}}^{(r)}\right)\boldsymbol{y} - \Delta^{(r)}\left(D\left(\widehat{w\rho_2}^{(r)}\right)1_n - \Delta^{(r)}D\left(\widehat{\rho_2^2}^{(r)}\right)\right)\boldsymbol{y}\right). \end{aligned}$$

(2) Update $\sigma^{2(r)}$ by

$$\sigma^{2(r+1)} = \frac{1}{n}\left(\boldsymbol{y} - \boldsymbol{X}\beta^{(r+1)}\right)^\top D\left(\widehat{\kappa_2^{-1}}^{(r)}\right)\left(\boldsymbol{y} - \boldsymbol{X}\beta^{(r+1)}\right).$$

(3) Update $\Delta^{(r)}$ by

$$\begin{aligned} \Delta^{(r+1)} = &\left(\left(\boldsymbol{y} - \boldsymbol{X}\beta^{(r+1)}\right)^\top D\left(\widehat{\rho_2^2}^{(r)}\right)\left(\boldsymbol{y} - \boldsymbol{X}\beta^{(r+1)}\right)\right)^{-1} \\ &\left(1_n^\top D\left(\widehat{w\rho_2}^{(r)}\right)\left(\boldsymbol{y} - \boldsymbol{X}\beta^{(r+1)}\right)\right), \end{aligned}$$

and then $\lambda^{(r+1)} = \sigma^{(r+1)}\Delta^{(r+1)}$.

(4) Update $\nu^{(r)}$ by

$$\nu^{(r+1)} = \text{argmax}_{\nu} \left[\sum_{i=1}^{n} E_2(\ln h(V_i)|y_i) \right].$$

(5) Update $\tau^{(r)}$ by

$$\tau^{(r+1)} = \text{argmax}_{\tau} \left[\sum_{i=1}^{n} E_2(\ln g(S_i)|y_i) \right].$$

In situations where M-steps (4 and 5) are challenging to compute, the following ML-steps can be used as alternatives.

(1) Update $\nu^{(r)}$ by

$$\nu^{(r+1)} = \text{argmax}_{\nu} \left[\sum_{i=1}^{n} \ln \int \phi(y_i; \boldsymbol{x}_i^\top \beta^{(r+1)}, \sigma^{2(r+1)} \kappa(v_i)) h(v_i) \mathrm{d}v_i \right].$$

(2) Update $\tau^{(r)}$ by

$$\tau^{(r+1)} = \text{argmax}_{\tau} \left[\sum_{i=1}^{n} \ln \int \Phi\left(\Delta^{(r+1)} \rho(s_i)(y_i - \boldsymbol{x}_i^\top \beta^{(r+1)}) \right) g(s_i) \mathrm{d}s_i \right].$$

The conditional expectations are computed as below:

$$E_2(\kappa^{-1}(V_i)|y_i) = \frac{\int \kappa^{-1}(v_i) \phi(y_i; \boldsymbol{x}_i^\top \beta, \sigma^2 \kappa(v_i)) h(v_i) \mathrm{d}v_i}{\int \phi(y_i; \boldsymbol{x}_i^\top \beta, \sigma^2 \kappa(v_i)) h(v_i) \mathrm{d}v_i},$$

$$E_2(\rho^2(S_i)|y_i) = \frac{\int \rho^2(s_i) \Phi(\Delta \rho(s_i)(y_i - \boldsymbol{x}_i^\top \beta)) g(s_i) \mathrm{d}s_i}{\int \Phi(\Delta \rho(s_i)(y_i - \boldsymbol{x}_i^\top \beta)) g(s_i) \mathrm{d}s_i},$$

$$E_2(W\rho(S_i)|y_i) = \Delta(y_i - \boldsymbol{x}_i^\top \beta) E_2(\rho^2(S_i)|y_i) + \frac{\int \rho(s_i) \phi(\Delta \rho(s_i)(y_i - \boldsymbol{x}_i^\top \beta)) g(s_i) \mathrm{d}s_i}{\int \Phi(\Delta \rho(s_i)(y_i - \boldsymbol{x}_i^\top \beta)) g(s_i) \mathrm{d}s_i}.$$

It is important to note that the reparameterization $\Delta = \frac{\lambda}{\sigma}$ is not intended to eliminate the dependence of the complete-data model on σ. Indeed, σ still appears in the Gaussian kernel through $\phi(y; x^\top \beta, \sigma^2, \kappa(\nu))$. Therefore, this transformation does not imply full separation between the scale and skewness parameters. Its main role is instead to provide a more convenient formulation for the EM algorithm. In particular, under the second scenario, the number of conditional expectations required in the E-step is reduced relative to the first scenario, yielding a more tractable updating scheme. Consequently, although both scenarios lead to essentially the same maximum-likelihood estimates, the reparametrized formulation may improve computational efficiency and facilitate subsequent theoretical developments, such as the derivation of the observed information matrix and diagnostic measures in more complex skew models.

Note: In this study, the log-likelihood function is employed as the primary convergence criterion; the EM algorithm terminates when the difference between the log-likelihood values at iterations $r+1$ and r falls below a threshold of 10^{-6}. Given that complex models such as the SHSSMN-LRM often require a large number of iterations to achieve convergence due to the inclusion of additional parameters (particularly for parameters such as skewness), we introduce an acceleration factor to enhance computational efficiency.
After a certain number of iterations (i.e., $r = 300$), if convergence has not been reached and the absolute difference between estimates for any component θ_i of the parameter vector θ falls below a predefined threshold (i.e., $\left|\theta_i^{(r+1)} - \theta_i^{(r)}\right| < 10^{-m_0}$ with $m_0 = 3$), the parameter estimate $\theta_i^{(r+1)}$ is updated as:

$$\theta_i^{(r)} + a\left(\theta_i^{(r+1)} - \theta_i^{(r)}\right),$$

where the acceleration coefficient is defined as $a = 10^{|m+m_0|}$, with $m = \log_{10}\left|\theta_i^{(r+1)} - \theta_i^{(r)}\right|$. This acceleration is applied subject to a likelihood-based safeguard: the update is accepted if and only if it results in an increase in the log-likelihood. If the accelerated step leads to a decrease in the log-likelihood, it is rejected, and the algorithm reverts to the standard EM update. Furthermore, if such a rejection occurs, m_0 can be adaptively increased (while maintaining $0 < m_0 < -m$) to ensure a more conservative step in subsequent iterations. This mechanism ensures that the acceleration provides significant computational savings without inducing instability or premature termination, effectively acting as an automated line-search in the neighborhood of the optimum.
We now proceed to compare the outcomes of the first and second scenarios:

1. In the second scenario, the number of conditional expectations required is fewer compared to the first scenario.

2. While the log-likelihood function remains identical in both scenarios, the complete log-likelihood functions differ.

3. The following common relationships between the conditional expectations can be identified:

$$\begin{aligned} E_1(\kappa^{-1}(V_i)|y_i) &= E_2(\kappa^{-1}(V_i)|y_i), \\ E_1(\rho^2(S_i)|y_i) &= E_2(\rho^2(S_i)|y_i), \\ E_1(W_i\rho(S_i)|y_i) &= \sigma E_2(W_i\rho(S_i)|y_i), \\ E_1(W_i^2|y_i) &= \sigma^2 E_2(W_i^2|y_i). \end{aligned}$$

3.2 Observed information matrix

Under some regularity conditions, the asymptotic covariance matrix of the ML estimates $\hat{\theta} = (\hat{\beta}^\top, \hat{\sigma}^2, \hat{\lambda})^\top$ can be approximated by the inverse of the observed information

matrix. We employ the method suggested by Louis (1982); Meilijson (1989). The empirical information matrix for the j-th scenario is defined as

$$I_j(\theta|\boldsymbol{y}) = \sum_{i=1}^{n} \widehat{\mathbb{S}_{ji}}\widehat{\mathbb{S}_{ji}}^{\top}, \qquad j = 1,2, \quad i = 1,2,\ldots,n,$$

where $\widehat{\mathbb{S}_{ji}} = (\widehat{\mathbb{S}_{ji,\beta}}^{\top}, \widehat{\mathbb{S}_{ji,\sigma^2}}, \widehat{\mathbb{S}_{ji,\lambda}})^{\top} = \frac{\partial Q_{ji}}{\partial \theta}\Big|_{\theta=\hat{\theta}}$, is the score statistic.The explicit expression for the elements of $\widehat{\mathbb{S}_{ji}}$ are summarized below:

$$\begin{aligned}
\widehat{\mathbb{S}_{1i,\beta}} &= \frac{1}{\hat{\sigma}^2}\{\hat{h}_{1i}(y_i - \boldsymbol{x}_i^{\top}\hat{\beta}) - \hat{\lambda}\widehat{w\rho_{1i}}\}\boldsymbol{x}_i,\\
\widehat{\mathbb{S}_{1i,\sigma^2}} &= -\frac{1}{\hat{\sigma}^2} + \frac{1}{2\hat{\sigma}^4}\{\hat{h}_{1i}(y_i - \boldsymbol{x}_i^{\top}\hat{\beta})^2 + \widehat{w_{1i}^2} - 2\hat{\lambda}\widehat{w\rho_{1i}}(y_i - \boldsymbol{x}_i^{\top}\hat{\beta})\},\\
\widehat{\mathbb{S}_{1i,\lambda}} &= \frac{1}{\hat{\sigma}^2}(y_i - \boldsymbol{x}_i^{\top}\hat{\beta})\{\widehat{w\rho_{1i}} - \hat{\lambda}\widehat{\rho_{1i}^2}(y_i - \boldsymbol{x}_i^{\top}\hat{\beta})\},\\
\widehat{\mathbb{S}_{2i,\beta}} &= \{\hat{h}_{2i}(y_i - \boldsymbol{x}_i^{\top}\hat{\beta}) - \hat{\Delta}\widehat{w\rho_{2i}}\}x_i,\\
\widehat{\mathbb{S}_{2i,\sigma^2}} &= -\frac{1}{\hat{\sigma}^2} + \frac{1}{2\hat{\sigma}^4}\widehat{\kappa_{2i}^{-1}}(y_i - \boldsymbol{x}_i^{\top}\hat{\beta})^2,\\
\widehat{\mathbb{S}_{2i,\lambda}} &= \frac{1}{\hat{\sigma}}\widehat{\mathbb{S}_{2i,\Delta}},
\end{aligned}$$

where

$$\widehat{\mathbb{S}_{2i,\Delta}} = (y_i - \boldsymbol{x}_i^{\top}\hat{\beta})\{\widehat{w\rho_{2i}} - \hat{\Delta}\widehat{\rho_{2i}^2}(y_i - \boldsymbol{x}_i^{\top}\hat{\beta})\}.$$

4 Simulation and Real Data Analysis

This section explores the influence of reparametrization on the EM algorithm and demonstrates the versatility of SHSSMN-LRM in various contexts. We first conduct a simulation study to evaluate the performance of parameters estimation under controlled scenarios. Our simulation design is intentionally within-family. Its main purpose is to investigate how reparametrization affects EM-based estimation in SHSSMN-LRM, both in terms of statistical accuracy and computational efficiency. For this reason, the simulation reports the average estimates, standard deviations, log-likelihood values, AIC, and the number of EM iterations across representative SHSSMN submodels. This design directly addresses the central methodological question of the paper, namely whether the reparametrized formulation achieves comparable inference with fewer iterations, especially for more complex distributions. Next, we apply the methodology to a real dataset to showcase the practical applicability and ability of SHSSMN-LRM to capture skewed and heavy-tailed data structures. All analyses and simulations were implemented in the R 4.3.3, R.Core.Team (2021) programming environment.

4.1 Simulation

Here, a simulation study is conducted on the LRM where the error term follows the SHSSMN family of distributions. The study aims to evaluate the performance of

the estimation method and analyze the influence of the two proposed scenarios on parameters estimation. To do so, consider the following linear regression model:

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i, \qquad i = 1, 2, \ldots, n,$$

where $\varepsilon_i \sim \text{STC}(0, \sigma^2, \lambda, \nu)$. The following combinations were taken for this simulation: $n = 100, 300, 1000$; $x_{ti} \sim \text{N}(0.1)$ for $t = 1, 2$; $\beta_1 = 1, \beta_2 = 2, \sigma^2 = 0.5, \lambda = -3, \nu = 1.5$. We generated 300 Monte Carlo samples and fitted the LRM to each sample under the STC distribution (for all sample sizes) and several other distributions from the SHSSMN family (only for $n = 1000$).

The initial values for the parameters $\beta^{(0)}$and $\sigma^{2(0)}$ were computed using the least squares method, while the initial value for $\lambda^{(0)}$ was determined from the skewness of the residuals. The stopping criterion for the algorithm was set as $|l(\theta^{(r+1)}) - l(\theta^{(r)})| < 10^{-6}$, where $l(\theta^{(r)})$ denotes the log-likelihood at the r-th iteration.

Table 1 presents the average values (Means) and standard deviations (SD's), the log-likelihood function (l) and the number of iterations (K) for the parameters of the STC-LRM for all sample sizes under both scenarios. It is evident that, in both scenarios, the average parameter estimates are close to the true values. As the sample size increases, both the mean and the SD of the ML estimators tend toward zero, confirming their consistency. Moreover, the outcomes for parameter estimation in each scenario are nearly identical. Table 2 shows l, AIC and K of fitting the LRM under various distributions from the SHSSMN family to simulated data with $n = 1000$. The table is used to identify the optimal model and compare the two proposed scenarios. The average reduction percentage (ARP) index was employed in each model to compare the number of iterations across scenarios for all simulation samples. The index is defined as follows:

$$\text{ARP} = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{K_{1j} - K_{2j}}{K_{1j}} \right) \times 100,$$

where N is the number of random simulation samples and $K_{ij}, i = 1, 2$ is the number of iterations for the scenario i and the j-th sample.

Based on Table 2, the model selection criteria for both scenarios are almost identical. Additionally, in the second scenario, the mean and SD of the K is fewer than in the first scenario, indicating a more efficient scenario to reach convergence. Notably, the ARP values show the second scenario for complicated distributions decrease the number of required iterations significantly and hence is more efficient. Table 2 presents the SD of the parameter estimates, l and K. It is apparent that, across both scenarios, the standard deviations of the parameter estimates are notably similar, indicating consistency in the model's performance under varying conditions.

4.2 A real data application

In this subsection, we analyses a real data set from Cook and Weisberg (2009) on characteristics of Australian athletes available from the Australian Institute of Sport

Table 1: Mean and SD for each estimator of STC-LRM under both scenarios.

Parameter	Scenario	$n = 100$		$n = 300$		$n = 1000$	
		Mean	SD	Mean	SD	Mean	SD
$\beta_1 = 1$	1	0.9961	0.0575	1.0005	0.0411	1.0011	0.0255
	2	0.9958	0.0574	1.0005	0.0410	1.0010	0.0255
$\beta_2 = 2$	1	2.0045	0.0671	1.9928	0.0412	1.9978	0.0216
	2	2.0041	0.0669	1.9928	0.0411	1.9978	0.0216
$\sigma^2 = 0.5$	1	0.5373	0.1631	0.5084	0.0897	0.4914	0.0522
	2	0.5371	0.1632	0.5080	0.0897	0.4909	0.0521
$\lambda = -3$	1	-3.7994	0.9333	-3.2536	0.7322	-3.0051	0.3241
	2	-3.7986	0.9331	-3.2442	0.7300	-2.9917	0.3277
$\nu = 1.5$	1	1.6086	0.3558	1.5194	0.1843	1.4800	0.1098
	2	1.6084	0.3562	1.5189	0.1843	1.4795	0.1096
l	1	-146.6713	14.8996	-443.0093	23.0026	-1487.9130	44.7313
	2	-146.6717	14.8996	-443.0094	23.0025	-1487.9141	44.7311
K	1	462.25	605.83	522.60	386.94	582.57	367.82
	2	365.77	343.08	354.00	175.63	333.30	127.59

(AIS). From the available variables, we select the three variables and apply the following LRM:

$$\text{LBM} = \beta_1 \text{WT} + \beta_2 \text{HT} + \varepsilon,$$

where LBM represents the lean body mass (kg); WT and HT are independent variables and denote respectively weight (kg) and height (cm); ε is an error term following the SHSSMN distribution. This data set was also studied by Arellano-Valle et al. (2008); Alizadeh Ghajari et al. (2024).

First, we fitted the Normal-LRM on the AIS data and the ML estimates were obtained. Table 3 provides the summary results for the ordinary residuals obtained from fitting the Normal-LRM. The skewness value of -1.02 indicates significant asymmetry. Also, the excess kurtosis (defined as the fourth standardized moment minus 3) value of 2.76 indicates that the distribution has fatter tails than the normal distribution (leptokurtic), consistent with the long-tailed residuals observed in Figure 1. Consequently, the Normal-LRM is deemed unsuitable for the AIS data, providing a robust empirical motivation for employing the SHSSMN-LRM, which is capable of capturing both heavy tails and asymmetry.

Table 2: Results of fitting the LRM under various distributions of the SHSSMN family on simulated data from model .

Distribution	Scenario	K			Model selection criterion			
					l		AIC	
		Mean	SD	ARP	Mean	SD	Mean	SD
SN	1	170.50	43.27	4.27	-2583.37	481.96	5174.75	963.92
	2	163.50	42.93		-2583.37	481.96	5174.74	963.92
SNC	1	338.63	131.65	1.54	-2544.28	486.31	5096.56	972.62
	2	332.80	128.96		-2544.28	486.31	5096.56	972.62
SGN	1	298.63	146.18	0.98	-2544.77	486.64	5099.54	973.28
	2	295.30	144.43		-2544.77	486.64	5099.54	973.28
STN	1	238.10	52.11	31.63	-1527.10	46.73	3064.20	93.46
	2	164.10	42.74		-1527.10	46.73	3064.20	93.46
SSLN	1	261.37	55.69	37.81	-1529.39	47.14	3068.77	94.27
	2	163.97	42.51		-1529.39	47.14	3068.77	94.27
MSTN	1	944.63	457.90	40.31	-1519.65	45.37	3049.31	90.74
	2	502.63	204.94		-1519.66	45.37	3049.33	90.75
MSSLN	1	981.50	522.68	43.54	-1522.04	45.71	3054.08	91.43
	2	492.73	205.28		-1521.98	45.82	3053.95	91.64
SSLC	1	409.00	125.19	29.57	-1490.23	45.13	2990.46	90.26
	2	285.23	116.45		-1490.23	45.13	2990.46	90.26
STC	1	582.57	367.82	34.04	-1487.91	44.73	2985.83	89.46
	2	333.30	127.59		-1487.91	44.7311	2985.82	89.46

The ML estimates are considered as the initial values for $\beta^{(0)}$ and $\sigma^{2(0)}$, while $\lambda^{(0)}$ is computed based on the skewness of the ordinary residuals. The convergence criterion employed is $|l(\theta^{(r+1)}) - l(\theta^{(r)})| < 10^{-6}$. In the proposed model, the SN, SNC, SGN, STN, SSLN, MSTN, MSSLN, SSLC, and STC distributions from the SHSSMN family were utilized to represent the error term distribution. Also, To evaluate the practical benefit of our proposed SHSSMN-LRM, we consider the Normal-LRM as a baseline reference. Table 4 reports the MLEs for the parameters, K and model selection criteria

Table 3: Summary results for residuals of Normal-LRM fitted on the AIS data .

Mean	SD	Median	Min	Max	Skewness	Kurtosis	K-S test statistic	p-value
0	2.27	0.24	-7.51	6.31	-1.02	2.76	0.129	0.000

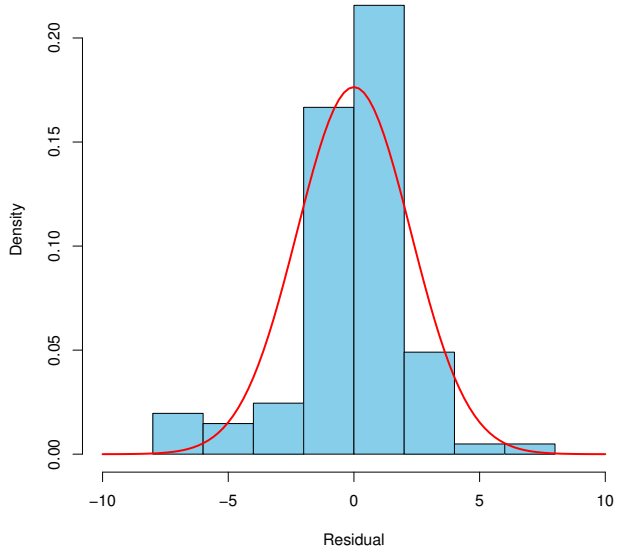

Figure 1: Histogram of Normal-LRM residuals with the fitted normal density curve .

including l and AIC for all SHSSMN-LRM members across both scenarios, alongside the Normal-LRM.

It is apparent that the parameter estimates and model selection criteria are remarkably similar across both scenarios for any error term distribution. Notably, for each error distribution, K is lower in the second scenario compared to the first, with more substantial reductions observed for the more complex distributions. Furthermore, all SHSSMN-LRM members exhibit superior fit compared to the Normal-LRM. In particular, the STC-LRM and SSLC-LRM yield the highest log-likelihood and lowest AIC, providing strong empirical evidence for their superior performance among all competing models. This is further substantiated by Figure 2, which depicts the histogram of the residuals from the STC-LRM, overlaid with the fitted STC density curve. The close alignment of the fitted STC distribution with the STC-LRM residuals suggests that the STC model effectively captures the underlying structure of the AIS dataset, particularly in accommodating asymmetry and heavy-tailed behavior.

Table 4: Results of fitting the SHSSMN-LRMs and the Normal-LRM baseline to the AIS dataset. The t-values are in parentheses

Distribution	Scenario	K	Parameter						Model selection criterion	
			β_1	β_2	σ^2	λ	τ	ν	l	AIC
Normal (baseline)	-	-	0.7317 (36.22)	0.0770 (7.62))	5.1140 (9.45)	-	-	-	-227.9631	461.9263
SN	1	249	0.8214 (45.16)	0.0509 (5.91)	12.5889 (7.20)	-4.6098 (-3.05)	-	-	-218.0179	444.03587
	2	234	0.8222 (45.52)	0.0506 (5.91)	12.6386 (7.24)	-4.6690 (-3.54)	-	-	-218.0173	444.03457
SNC	1	513	0.7780 (32.39)	0.0685 (6.17)	10.3386 (7.99)	-17.6380 (-1.86)	-	-	-213.6069	435.2138
	2	496	0.7780 (32.37)	0.0685 (6.16)	10.3428 (7.99)	17.6407 (-1.94)	-	-	-213.6069	435.2138
SGN	1	417	0.7733 (30.43)	0.0707 (6.07)	10.4572 (7.52)	-10.2394 (-1.75)	0.2583	-	213.6184	437.2369
	2	398	0.7734 (30.42)	0.0707 (6.07)	10.4636 (7.53)	-10.2627 (-1.84)	0.2583	-	-213.6186	437.2371
STN	1	274	0.8122 (48.20)	0.0481 (7.04)	1.3757 (3.06)	-0.5381 (-1.50)	-	1.7285	208.6829	427.3659
	2	182	0.8111 (47.96)	0.0484 (7.05)	1.3475 (3.12)	-0.5081 (-1.78)	-	1.7244	-208.6811	427.3621
SSLN	1	374	0.8090 (48.26)	0.0489 (7.17)	0.5721 (3.25)	-0.2931 (-1.36)	-	0.6711	-209.0492	428.0985
	2	245	0.8090 (48.27)	0.0489 (7.17)	0.5711 (3.25)	-0.2928 (-1.36)	-	0.6706	-209.0492	428.0985
MSTN	1	568	0.7913 (53.61)	0.0566 (6.96)	1.4138 (1.84)	-0.6476 (-0.36)	-	1.8493	-210.2535	430.5069
	2	388	0.7913 (53.71)	0.0565 (6.79)	1.3973 (1.74)	-0.6227 (0.36)	-	1.8431	-210.2523	430.5047
MSSLN	1	722	0.7880 (54.77)	0.0576 (6.31)	0.6457 (1.57)	-0.3664 (-0.28)	-	0.7237	-210.5249	431.0499
	2	528	0.7880 (54.87)	0.0575 (5.81)	0.6356 (1.36)	-0.3447 (-0.25)	-	0.7210	-210.5231	431.0462
SSLC	1	310	0.7996 (48.89)	0.0584 (8.26)	2.4618 (4.11)	-8.9949 (-1.71)	-	1.0589	-207.2130	424.4259
	2	258	0.7995 (48.93)	0.0584 (8.28)	2.4762 (4.13)	-9.1596 (-1.88)	-	1.0622	-207.2122	424.4245
STC	1	364	0.8010 (47.99)	0.0577 (8.05)	4.4041 (4.13)	-11.8558 (-1.69)	-	3.0860	-207.0781	424.1563
	2	256	0.8009 (47.99)	0.0579 (8.05)	4.4187 (4.14)	-11.9799 (-1.86)	-	3.0945	-207.0778	424.1555

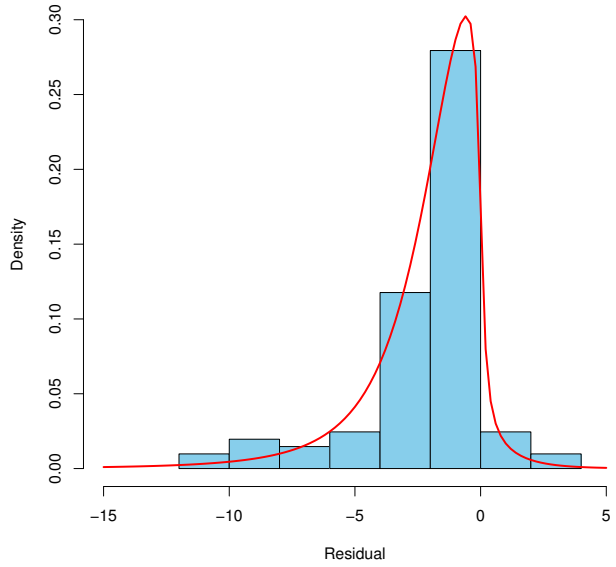

Figure 2: Histogram of STC-LRM residuals with the fitted STC density curve.

5 Discussion

We proposed a novel and flexible class of linear regression models designed for cases where the error term follows the SHSSMN distribution. These models provide a more comprehensive framework compared to earlier studies. To estimate the model parameters, we extended the EM algorithm under two distinct scenarios based on without or with reparametrization. In the second scenario, the number of conditional expectations is reduced compared to the first. Therefore, as seen in Section 4, the reduction in the number of iterations required for convergence in the second scenario suggests a computational advantage, making it a more efficient approach for parameter estimation in complex distributional settings.

The results from our simulation study highlight the efficacy of the proposed models in capturing the underlying structure of skewed and heavy-tailed data. Both the first and second scenarios yield nearly identical results in terms of parameter estimates, demonstrating the robustness and consistency of the EM algorithm across different configurations. However, the second scenario (the model under reparameterization) has the advantage of requiring fewer iterations to converge, particularly when dealing with complex error distributions. This reduction in computational effort suggests that the second scenario may be a more efficient choice for large-scale or computationally intensive applications.

Moreover, the real-world application of the SHSSMN models to the AIS dataset re-

inforces their practical utility. The models effectively address the limitations of the traditional Normal-LRM by accounting for skewness and heavy tails in the residuals. The STC distribution, in particular, provides the best fit according to model selection criteria such as log-likelihood and AIC. This outcome demonstrates the model's ability to adapt to the specific data characteristics of the AIS dataset, making it a valuable tool for similar applications in other fields.
Overall, the present work opens several avenues for future investigation. In particular, the reparameterized SHSSMN framework may be extended to other statistical models, while the associated estimation procedures may be further refined and complemented with additional optimization strategies to improve computational efficiency. In addition, a broader comparative analysis of SHSSMN-LRM relative to classical competing models would be worthwhile, although this lies outside the scope of the current paper and is left for future research.

References

Alizadeh Ghajari Z, Zare K, Shokri S. Estimation in shape mixtures of skew-normal linear regression models via ECM coupled with Gibbs sampling. Monte Carlo Methods and Applications. 2024;30(2):137–148.

Arellano-Valle RB, Castro LM, Genton MG, Gómez HW. Bayesian inference for shape mixtures of skewed distributions, with application to regression analysis. 2008;.

Arellano-Valle RB, Ferreira CS, Genton MG. Scale and shape mixtures of multivariate skew-normal distributions. Journal of Multivariate Analysis. 2018;166:98–110.

Arellano-Valle RB, Genton MG, Loschi RH. Shape mixtures of multivariate skew-normal distributions. Journal of Multivariate Analysis. 2009;100(1):91–101.

Arellano-Valle RB, Gómez HW, Quintana FA. A new class of skew-normal distributions. Communications in statistics-Theory and Methods. 2004;33(7):1465–1480.

Arnold BC, Beaver RJ. The skew-Cauchy distribution. Statistics & probability letters. 2000;49(3):285–290.

Ashour SK, Abdel-hameed MA. Approximate skew normal distribution. Journal of Advanced Research. 2010;1(4):341–350.

Azzalini A. A class of distributions which includes the normal ones. Scandinavian journal of statistics. 1985;p. 171–178.

Azzalini A, Capitanio A. Statistical applications of the multivariate skew normal distribution. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 1999;61(3):579–602.

Azzalini A, Capitanio A. Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution. Journal of the Royal Statistical Society Series B: Statistical Methodology. 2003;65(2):367–389.

Branco MD, Dey DK. A general class of multivariate skew-elliptical distributions. Journal of Multivariate Analysis. 2001;79(1):99–113.

Celeux G. The SEM algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. Computational statistics quarterly. 1985;2:73–82.

Cook RD, Weisberg S. An introduction to regression graphics, vol. 405. John Wiley & Sons; 2009.

Delyon B, Lavielle M, Moulines E. Convergence of a stochastic approximation version of the EM algorithm. Annals of statistics. 1999;p. 94–128.

Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. Journal of the royal statistical society: series B (methodological). 1977;39(1):1–22.

Ferreira C. Diagnostics Analysis for Skew Normal Regression Models. Instituto de Matemática, Estatística e Computação Científica, Universidade . . . ; 2008.

Ferreira C, Bolfarine H, Lachos VH. Skew scale mixtures of normal distributions: Properties and estimation. Statistical Methodology. 2011;8(2):154–171.

Ferreira CS, Lachos VH, Bolfarine H. Inference and diagnostics in skew scale mixtures of normal regression models. Journal of Statistical Computation and Simulation. 2015;85(3):517–537.

Genton MG, Loperfido NM. Generalized skew-elliptical distributions and their quadratic forms. Annals of the Institute of Statistical Mathematics. 2005;57(2):389–401.

Gómez HW, Venegas O, Bolfarine H. Skew-symmetric distributions generated by the distribution function of the normal distribution. Environmetrics: The official journal of the International Environmetrics Society. 2007;18(4):395–407.

Gupta AK, Chang F, Huang WJ. Some skew-symmetric models. Random Operators & Stochastic Equations. 2002;10(2).

Kahrari F, Rezaei M, Yousefzadeh F, Arellano-Valle RB. On the multivariate skew-normal-Cauchy distribution. Statistics & Probability Letters. 2016;117:80–88.

Kuhn E, Lavielle M. Coupling a stochastic approximation version of EMwith an MCMC procedure. ESAIM: Probability and Statistics. 2004;8:115–131.

Louis TA. Finding the observed information matrix when using the EM algorithm. Journal of the Royal Statistical Society Series B: Statistical Methodology. 1982;44(2):226–233.

Mattos TdB, Garay AM, Lachos VH. Likelihood-based inference for censored linear regression models with scale mixtures of skew-normal distributions. Journal of Applied Statistics. 2018;45(11):2039–2066.

Meilijson I. A fast improvement to the EM algorithm on its own terms. Journal of the Royal Statistical Society Series B: Statistical Methodology. 1989;51(1):127–138.

Meng XL, Rubin DB. Maximum likelihood estimation via the ECM algorithm: A general framework. Biometrika. 1993;80(2):267–278.

Nadarajah S, Kotz S. Skewed distributions generated by the normal kernel. Statistics & probability letters. 2003;65(3):269–277.

Nematollahi N, Farnoosh R, Rahnamaei Z. Location-scale mixture of skew-elliptical distributions: Looking at the robust modeling. Statistical Methodology. 2016;32:131–146.

Team RC. R: A language and environment for statistical computing. R foundation for statistical computing, Vienna, Austria. 2021;.

Sahu SK, Dey DK, Branco MD. A new class of multivariate skew distributions with applications to Bayesian regression models. Canadian Journal of Statistics. 2003;31(2):129–150.

Vilca F, Balakrishnan N, Zeller CB. Multivariate skew-normal generalized hyperbolic distribution and its properties. Journal of Multivariate Analysis. 2014;128:73–85.

Wang J, Genton MG. The multivariate skew-slash distribution. Journal of Statistical Planning and Inference. 2006;136(1):209–220.

Wang J, Boyer J, Genton MG. A skew-symmetric representation of multivariate distributions. Statistica Sinica. 2004;p. 1259–1270.

Watagoda L, Don HSRA, Sanqui JAT. A cosine approximation to the skew normal distribution. In: Int. Math. Forum, vol. 14; 2019. p. 253–261.

Wei GC, Tanner MA. A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. Journal of the American statistical Association. 1990;85(411):699–704.