JIRSS (2026)

Vol. 25, No. 2, pp 79-118

DOI: 10.22034/jirss.2026.2081234.1170

Expert mixture models based on extended skew-normal distributions for financial data analysis

Farzane Hashemi

Department of Statistics, Faculty of Mathematical Sciences, University of Kashan, Kashan, Iran.

Received: 2025-12-18, Accepted: 2026-07-26, Published online: 2026-08-12

Abstract. The mixture-of-experts (MoE) framework provides a flexible probabilistic approach for modeling, classification, and clustering of heterogeneous data by combining several local expert models under a gating mechanism. Each expert specializes in a subset of the input space, while the gating function determines the data-to-expert allocation through covariate-dependent mixing proportions. Although the classical MoE formulation-typically assuming normally distributed error terms-offers analytical convenience, it often lacks robustness when data exhibit asymmetry, heavy tails, or contamination. To overcome these challenges, we propose a novel robust non-normal MoE model in which each expert component follows a log scale-shape mixture of skew-normal (SSMSN) distribution. This formulation, belonging to the broader SSMSN family, allows the model to flexibly capture skewed and heavy-tailed behaviors that frequently arise in economic and financial datasets. To efficiently estimate the model parameters, we propose an Expectation-Maximization (EM) algorithm that maximizes the likelihood function to obtain the maximum likelihood (ML) estimates. We conduct extensive Monte Carlo simulations to assess the finite-sample properties, robustness, and accuracy of the proposed estimator. Finally, the practical usefulness of the proposed MoE-SSMSN framework is demonstrated through an empirical applica-

Corresponding Author: Farzane Hashemi (farzane.hashemi1367@kashanu.ac.ir).

tion to real data analysis, highlighting its superiority in modeling heterogeneous, and skewed observations.

Keywords. Mixture of experts, Scale-shape mixture of skew normal, Robust estimation, Heavy-tailed data, Financial modeling.

MSC: 62-XX, 62Rxx, 62R10.

1 Introduction

Mixture-of-experts (MoE) models provide a powerful statistical and machine-learning framework for modeling, classification, and clustering of heterogeneous data by combining multiple local regression models under a gating mechanism (Jacobs et al., 1991; Nguyen and McLachlan, 2016; Chamroukhi, 2017). Each expert specializes in a subregion of the input space, while the gating function determines the probabilistic assignment of observations to experts through covariate-dependent mixing proportions. Due to their flexibility and interpretability, MoE models have been successfully applied in a wide range of fields, including economics, marketing, bioinformatics, and finance (Mazza and Punzo, 2017; Wang et al., 1998; Gormley and Murphy, 2008).

A common assumption in classical MoE formulations is that the expert error terms follow a Gaussian distribution. Although this assumption leads to computationally convenient inference via expectation–maximization (EM) algorithms (Dempster et al., 1977), it often lacks robustness in the presence of skewness, heavy tails, or outliers. Such features are frequently encountered in economic and financial data, where observations may arise from multiple latent sources of heterogeneity, leading to asymmetric and heavy-tailed behavior. Inference based on symmetric distributions in these settings can result in biased parameter estimates and poor clustering or prediction performance (Zeller et al., 2016). Consequently, extending MoE models beyond the Gaussian assumption has become an active area of research.

To address asymmetry, several authors have incorporated skewed distributions into mixture regression and MoE frameworks, most notably the skew-normal distribution originally introduced by Azzalini (1985, 1986), along with its heavy-tailed extensions such as the skew-t distribution. The extensive literature on skew-elliptical distributions initiated by Azzalini and co-authors (Azzalini and Dalla Valle, 1996; Azzalini

and Capitanio, 2003) has motivated numerous robust extensions of mixture and regression models. Further developments include Laplace, contaminated normal, and other robust distributions (Nguyen and McLachlan, 2016; Chamroukhi, 2017; Mazza and Punzo, 2020). While these models improve flexibility relative to Gaussian-based approaches, they may still be inadequate for capturing pronounced right-skewness and extreme tail behavior commonly observed in practice.

To overcome these limitations, we propose a new robust MoE framework in which each expert component follows a scale-shape mixture of skew-normal (SSMSN) distribution (Jamalizadeh and Lin, 2017). The SSMSN family extends the classical skew-normal distribution of Azzalini (1985, 1986) by introducing latent scale and shape mixing variables, thereby enabling simultaneous modeling of skewness and heavy-tailed behavior. While the skew-normal distribution introduced by Azzalini preserves many desirable properties of the normal distribution, its tail behavior remains Gaussian; the proposed SSMSN model overcomes this limitation by augmenting the skew-normal kernel with latent scale and shape mixing. Importantly, the proposed family includes the standard skew-normal distribution as a special case, obtained when the scale and shape mixing components degenerate to fixed constants. Although the SSMSN distribution does not belong to the conventional scale-mixture-of-normals (SMN) class, it admits a tractable hierarchical latent-variable representation that is particularly well suited for EM-type inference.

The remainder of this paper is organized as follows. Section 2 briefly introduces the SSMSN distribution and discusses its key properties. Section 3 formulates the MoE-SSMSN model and describes the EM-type algorithm for parameter estimation. Section 4 outlines the derivation of theoretical formulas for value-at-risk and tail value-at-risk using the MoE-SSMSN distributions. Section 5 presents simulation studies to assess the finite-sample performance and robustness of the proposed method. Section 6 illustrates the application of the model to real economics data. Finally, Section 7 concludes the paper with some remarks and directions for future research.

2 The family of SSMSN distributions

This section offers a concise overview of the SSMSN distribution family, which will be used later in the development of the MoE models. Consider two independent random variables Z_1 and Z_2, both of which follow the standard normal distribution,

i.e., $Z_1, Z_2 \sim \mathrm{N}(0,1)$. Let $\boldsymbol{\tau} = (\tau_1, \tau_2)^\top$ be a random vector, where both components are positive. The joint distribution of $\boldsymbol{\tau}$ is given by the probability density function $h(\boldsymbol{\tau}; \boldsymbol{\nu})$, with $\boldsymbol{\nu}$ representing the parameters of the distribution. Additionally, assume that $(Z_1, Z_2)^\top$ and $\boldsymbol{\tau}$ are independent. A random variable Y is said to follow the SSMSN distribution, denoted by $Y \sim \mathrm{SSMSN}(\mu, \sigma^2, \lambda, \boldsymbol{\nu})$, if the conditional distribution of Y given $\boldsymbol{\tau} = (\tau_1, \tau_2)^\top$ is as follows:

$$Y \mid (\tau_1, \tau_2) \overset{d}{=} SN\left(\mu, \sigma^2 \tau_1^{-1}, \lambda \left(\frac{\tau_1}{\tau_2}\right)^{-1/2}\right), \tag{2.1}$$

where $\mathrm{SN}(\mu, \sigma^2, \lambda)$ refers to the skew-normal distribution with location μ, scale σ^2, and shape parameter λ.

By utilizing the stochastic representation given in equation (2.1), the random variable Y can be expressed as:

$$Y \overset{d}{=} \mu + \sigma \left(\frac{\tau_1^{-\frac{1}{2}} f^{\frac{1}{2}}}{\sqrt{f + \lambda^2}} Z_2 + \frac{\lambda \tau_1^{-\frac{1}{2}}}{\sqrt{f + \lambda^2}} |Z_1| \right), \tag{2.2}$$

with $f = \tau_1/\tau_2$. Alternatively, denoting $\gamma = \sqrt{\tau_1^{-1}(f + \lambda^2)}|Z_1|$, a practical hierarchical form for the SSMSN distribution can be represented as:

$$Y \mid \gamma, \tau_1, \tau_2 \sim N\left(\mu + \frac{\sigma\lambda\gamma}{f + \lambda^2}, \frac{\sigma^2 f}{\tau_1(f + \lambda^2)}\right), \quad \gamma \mid \tau_1, \tau_2 \sim TN\left(0, \frac{f + \lambda^2}{\tau_1}; (0, \infty)\right), \tag{2.3}$$

where $TN(\mu, \sigma^2; (a, b))$ denotes the truncated normal distribution with mean μ, variance σ^2, truncated to the interval (a, b); see Jamalizadeh and Lin (2017) for further details.

From Equations (2.3), the joint pdf of Y, γ, and $\boldsymbol{\tau} = (\tau_1, \tau_2)^\top$ is expressed as

$$\begin{aligned} f(y, \gamma, \boldsymbol{\tau}) = & \frac{1}{\sigma\pi} \exp\left\{-\frac{1}{2}u^2\tau_1 - \frac{1}{2}\gamma^2\tau_2 + \lambda u\gamma\tau_2 - \frac{1}{2}\lambda^2\tau_2 u^2\right\} \\ & \times \sqrt{\tau_1}\sqrt{\tau_2} h(\tau_1, \tau_2; \boldsymbol{\nu}), \end{aligned} \tag{2.4}$$

where $h(\boldsymbol{\tau}; \boldsymbol{\eta})$ denotes the pdf of the positive bivariate mixing variable $\boldsymbol{\tau} = (\tau_1, \tau_2)^\top$, and $u = \frac{y-\mu}{\sigma}$.

By integrating out the mixing variable γ in (2.4), the marginal pdf of Y and $\boldsymbol{\tau}$ is obtained as

$$f(y, \boldsymbol{\tau}) = \frac{2}{\sigma} \phi(\sqrt{\tau_1}\, u)\, \Phi\left(\lambda \sqrt{\tau_2}\, u\right) \sqrt{\tau_2}\, h(\tau_1, \tau_2; \boldsymbol{\nu}). \tag{2.5}$$

The pdf of Y is then derived by integrating out the mixing variables τ_1 and τ_2 following

$$f(y) = \int_0^\infty \int_0^\infty f(y, \boldsymbol{\tau})\, d\tau_1\, d\tau_2.$$

By dividing equation (2.4) by equation (2.5), the conditional density is acquire as:

$$f(\gamma \mid y, \tau_1, \tau_2) = \frac{\sqrt{\tau_2}\,\phi\left(\sqrt{\tau_2}(\gamma - u)\right)}{\Phi\left(\sqrt{\tau_2}\,u\right)} = f(\gamma \mid y, \tau_2), \tag{2.6}$$

which implies that given $Y = y$ and τ_2, the variables τ_1 and τ_2 are conditionally independent. From equation (2.6), we deduce that:

$$\gamma \mid (y, \tau_2) \sim TN\left(u, \frac{1}{\tau_2}, (0, \infty)\right),$$

where $\text{TN}(\mu, \sigma^2; (0, \infty))$ refers to the truncated normal distribution with mean μ, variance σ^2, and support on $(0, \infty)$.

The conditional pdf of $\boldsymbol{\tau} = (\tau_1, \tau_2)^\top$ given $Y = y$ can be expressed as

$$f(\tau_1, \tau_2 \mid y) = \frac{2\,h(\tau_1, \tau_2; \boldsymbol{\nu})\,\sqrt{\tau_1}\,\phi\,(\tau_1\,u)\,\Phi\left(\sqrt{\tau_2}\,u\right)}{f(t)}, \tag{2.7}$$

where $h(\tau_1, \tau_2; \boldsymbol{\nu})$ denotes the joint pdf of $\boldsymbol{\tau} = (\tau_1, \tau_2)^\top$, and $f(y)$ is the marginal pdf of Y.

Some special distributions inherited in the SSMSN family can be formed by varying the specification of $h(\boldsymbol{\tau}; \boldsymbol{\nu})$. More details on the SSMSN family formulation are sketched in AppendixA. Also in AppendixA, we give an overview on some particular cases of the SSMSN class of distributions. These particular members include the skew-*t*-*t* (STT), skew-*t* (ST), skew generalized Laplace normal (SGLN) and skew slash normal (SSN) distributions.

3 Mixture of SSMSN experts

In this section, we introduce a new class of regression models in which the error terms follow a SSMSN distributions.

3.1 Model formulation

To address nonlinear regression challenges and model-based clustering in the context of correlated responses, we adapt the mixture of experts regression (MoER) approach, enhancing its flexibility for these complex data scenarios. Let $Z \in \{1, \ldots, g\}$ be a

latent class variable indicating the component (or expert) membership. The mixing probabilities are modeled through a multinomial logistic function as

$$\pi_i(\boldsymbol{w};\boldsymbol{\psi}) = P(Z = i \mid \boldsymbol{w}) = \begin{cases} \dfrac{\exp(\boldsymbol{\psi}_j^\top \boldsymbol{w})}{1 + \sum_{k=1}^{g-1} \exp(\boldsymbol{\psi}_k^\top \boldsymbol{w})}, & \text{if } i = 1, \ldots, g-1, \\ \dfrac{1}{1 + \sum_{k=1}^{g-1} \exp(\boldsymbol{\psi}_k^\top \boldsymbol{w})}, & \text{if } i = g, \end{cases}$$

where $w \in \mathbb{R}^q$ represents a set of concomitant variables, while $\boldsymbol{\psi}_i \in \mathbb{R}^q$ denotes the parameter vector associated with the ith expert. Additionally, $\boldsymbol{\psi} = (\boldsymbol{\psi}_1^\top, \ldots, \boldsymbol{\psi}_{g-1}^\top)^\top$ corresponds to the complete parameter vector for the gating network.

Let $\boldsymbol{X} \in \mathbb{R}^p$ denote the covariate (feature) vector, composed of p input variables $(X_1, \ldots, X_p)^\top$, and let $Y \in \mathbb{R}$ denote the response variable. Conditional on $Z = i$, the MoER model assumes the following component-specific regression relationship:

$$Y = \boldsymbol{\varphi}_i^\top \tilde{\boldsymbol{x}} + \varepsilon_i, \quad i = 1, \ldots, g, \tag{3.1}$$

where $\tilde{\boldsymbol{x}} = (1, \boldsymbol{x}^\top)^\top$ is the augmented covariate vector (typically including an intercept term), $\boldsymbol{\varphi}_i = (\varphi_{0i}, \varphi_{1i}, \ldots, \varphi_{pi})$ is the $(p+1)$ vector of regression coefficients for component i, and ε_i denotes the error term assumed to follow a SSMSN distribution.

In order to develop a more stable version of the regression model, we hypothesize that the residuals adhere to the SSMSN distribution, expressed as:

$$\varepsilon_i \sim \text{SSMSN}(0, \sigma_i^2, \lambda_i, \boldsymbol{\nu}_i).$$

Assuming that $\boldsymbol{Z}$ is independent of $\boldsymbol{x}$ and $\boldsymbol{w}$. The conditional pdf of Y is formulated as:

$$f(y \mid \boldsymbol{x}, \boldsymbol{w}; \boldsymbol{\Theta}) = \sum_{i=1}^{g} \pi_i(\boldsymbol{w};\boldsymbol{\psi})\, f_{\text{SSMSN}}\Big(y; \boldsymbol{\varphi}_i^\top \tilde{\boldsymbol{x}}, \sigma_i^2, \lambda_i, \boldsymbol{\nu}_i\Big), \tag{3.2}$$

where $\pi_i(\boldsymbol{w};\boldsymbol{\psi})$, for $i = 1, \ldots, g$, represents the mixing proportions, also known as the gating functions in the MoE framework. The complete parameter vector is given by $\boldsymbol{\Theta} = \left(\boldsymbol{\psi}_1^\top, \ldots, \boldsymbol{\psi}_{g-1}^\top, \boldsymbol{\theta}_1^\top, \ldots, \boldsymbol{\theta}_g^\top\right)^\top$, where each expert-specific parameter vector is defined as $\boldsymbol{\theta}_j = \left(\boldsymbol{\varphi}_i, \sigma_i^2, \lambda_i, \boldsymbol{\nu}_i\right)^\top$. It is worth noting that the explanatory variables $\boldsymbol{x}$ and the concomitant variables $\boldsymbol{w}$ may be identical or partially overlapping, depending on the modeling context.

We will henceforth mention to the model expressed in (3.2) as the MoER-SSMSN model. The term “SSMSN" in the subscript signifies that the error components, or

the individual expert distributions, are modeled by the SSMSN distribution. It is essential to observe that when $\lambda_i = 0$ for all $i = 1, \ldots, g$, the model simplifies to the MoER-T model, where the expert distributions become symmetric and are based on the t-student distribution. Moreover, in the case where $\boldsymbol{w}$ uninfluence the gating function the model from (3.2) is simplified to a simpler form of mixture regression models, referred to as MR-SSMSN, where the mixing proportions remain constant across different observations.

With the maximum likelihood (ML) estimator for $\boldsymbol{\Theta}$, we can easily compute the mean of the MoER-SSMSN model. In the subsequent section, a method based on EM-type algorithm is proposed for estimating the parameters of the MoER-SSMSN model, with the convolution-based form of the SSMSN distribution being utilized.

3.2 Parameter Estimation

Consider a random sample $(y_1, \ldots, y_n)$, where the observations are independent and identically distributed (i.i.d.), with $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ and $(\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n)$, drawn from a distribution whose density is described by the expression in (3.2). The log-likelihood function for the parameter vector $\boldsymbol{\Theta}$, based on the observed data, is then given by: The observed data-based log-likelihood function for the unknown parameter $\boldsymbol{\Theta}$ is expressed as follows:

$$\ell(\boldsymbol{\Theta}; \boldsymbol{y}, \boldsymbol{x}, \boldsymbol{w}) = \sum_{j=1}^{n} \log \left(\sum_{i=1}^{g} \pi_i(\boldsymbol{w}; \psi) \, f_{\text{SSMSN}} \left(y; \boldsymbol{\varphi}_i^{\top} \tilde{\boldsymbol{x}}, \sigma_i^2, \lambda_i, \boldsymbol{\nu}_i \right) \right), \tag{3.3}$$

The log-likelihood function given in (3.3) is intricate, making it impossible to directly derive the maximum likelihood estimates (MLEs) for the model parameters in an explicit form. To overcome this challenge, we employ an Expectation-Maximization (EM)-based approach, specifically the Expectation Conditional Maximization (ECM) algorithm (Meng and Rubin, 1993), for parameter estimation. In certain cases, we further utilize the Expectation Conditional Maximization Either (ECME) algorithm (Liu and Rubin, 1994), where some of the ECM's conditional maximization steps are replaced by constrained maximization steps (CML), aimed at optimizing the corresponding constrained likelihood function.

Let $\boldsymbol{Z}_j = (Z_{1j}, \ldots, Z_{gj})^{\top}$ represent an auxiliary latent random vector, where each component Z_{ij} indicates the membership of the j-th observation to class i. Let Z_{ij} be an indicator variable, where $Z_{ij} = 1$ indicates that the observation belongs to class i,

and $Z_{ij} = 0$ otherwise. By utilizing the convolution-based representation of the SSMSN distribution, as outlined in (2.3), a comprehensive hierarchical model can be developed for the MoE-SSMSN framework, as shown in (3.2),

$$\begin{aligned} Y_j \mid \boldsymbol{x}_j, \gamma_j, \tau_{1j}, \tau_{2j}, Z_{ij} = 1 \sim & \mathrm{N}\left(\boldsymbol{\varphi}_i^\top \tilde{\boldsymbol{x}}_j + \frac{\sigma_i \lambda_i}{f_j + \lambda_i^2}\gamma_j, \frac{\sigma_i^2 f_j}{\tau_{1j}(f_j + \lambda_i^2)}\right), \\ \gamma_j \mid \tau_{1j}, \tau_{2j}, Z_{ij} = 1 \sim & TN\left(0, \frac{f_j + \lambda_i^2}{\tau_{1j}}\right) \\ \tau_{1j}, \tau_{2j} \mid Z_{ij} = 1 \sim & h(\tau_{1j}, \tau_{2j}; \boldsymbol{\nu}), \\ \boldsymbol{Z}_j \mid \boldsymbol{w}_j \sim & M(1; \pi_1(\boldsymbol{w}_j; \boldsymbol{\psi}), \ldots, \pi_g(v_j; \boldsymbol{\psi})), \quad j = 1, \ldots, n, \ i = 1, \ldots, g, \end{aligned} \tag{3.4}$$

where the variable $f_j = \frac{\tau_{1j}}{\tau_{2j}}$ and $\gamma_j = \sqrt{\tau_{1j}^{-1}(f_j + \lambda_i^2)}\ |Z_1|$, while $M(1; \cdot)$ represents the multinomial distribution in a single trial, with probabilities $\pi_1(\boldsymbol{w}_j; \boldsymbol{\psi}), \ldots, \pi_g(\boldsymbol{w}_j; \boldsymbol{\psi})$. It is important to emphasize that the triplet $(\gamma_j, \tau_{1j}, \tau_{2j})$ is regarded as a latent variable within this hierarchical model.

Consider $\boldsymbol{y}_c = (\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{w}, \boldsymbol{\gamma}, \boldsymbol{\tau}, \boldsymbol{z})$ indicate the complete set of data. Based on the hierarchical structure outlined in (3.4), the log-likelihood function for the complete data is given by

$$\begin{aligned} \ell_c(\Theta; y_c) = \sum_{j=1}^{n}\sum_{i=1}^{g} z_{ij}\Big[& \log \pi_i(\boldsymbol{w}_j; \boldsymbol{\psi}) - \frac{\tau_{1j}(y_j - \boldsymbol{\varphi}_i^\top \tilde{\boldsymbol{x}}_j)^2}{2\sigma_i^2} + \eta_i \tau_{2j}\gamma_j(y_j - \boldsymbol{\varphi}_i^\top \tilde{\boldsymbol{x}}_j) \\ & - \frac{\eta_i^2 \tau_{2j}(y_j - \boldsymbol{\varphi}_i^\top \tilde{\boldsymbol{x}}_j)^2}{2} - \log(\sigma_i) + \log\Big(\sqrt{\tau_{1j}\tau_{2j}} h(\tau_{1j}, \tau_{2j}; \boldsymbol{\nu}_i)\Big)\Big], \end{aligned} \tag{3.5}$$

where $\eta_i = \lambda_i / \sigma_i$.

According to the ECME algorithm framework, at the k^{th} iteration, the conditional expectation of the complete-data log-likelihood function in equation (3.5) must be computed, with the parameter vector $\boldsymbol{\Theta}$ substituted by its estimate at the k-th iteration, $\boldsymbol{\Theta}^{(k)}$

$$\begin{aligned} Q(\boldsymbol{\Theta} \mid \boldsymbol{\Theta}^{(k)}) = \sum_{j=1}^{n}\sum_{i=1}^{g} \hat{z}_{ij}^{(k)}\Big[& \log \pi_i(\boldsymbol{w}_j; \boldsymbol{\psi}) - \frac{(y_j - \boldsymbol{\varphi}_i^\top \tilde{\boldsymbol{x}}_j)^2 \widehat{W}_{1ij}^{(k)}}{2\sigma_i^2} + \eta_i(y_j - \boldsymbol{\varphi}_i^\top \tilde{\boldsymbol{x}}_j)\widehat{W}_{3ij}^{(k)} \\ & - \frac{\eta_i^2 (y_j - \boldsymbol{\varphi}_i^\top \tilde{\boldsymbol{x}}_j)^2 \widehat{W}_{2ij}^{(k)}}{2} - \log(\sigma_i) + \mathrm{E}\left(\log\left(\sqrt{\tau_{1j}\tau_{2j}} h(\tau_{1j}, \tau_{2j}; \boldsymbol{\nu}_i)\right) \mid y_j, \boldsymbol{\theta}_i^{(k)}\right)\Big], \end{aligned} \tag{3.6}$$

where

$$z_{ij}^{(k)} = \frac{\pi_i(\boldsymbol{w}_j;\hat{\boldsymbol{\psi}}^{(k)})\, f_{\text{SSMSN}}\left(y_j;\, \tilde{\boldsymbol{x}}_j^\top \hat{\boldsymbol{\varphi}}_i^{(k)},\, \hat{\sigma}_i^{2(k)},\, \hat{\lambda}_i^{(k)},\, \hat{\boldsymbol{\nu}}_i^{(k)}\right)}{\sum_{l=1}^{g} \pi_l(\boldsymbol{w}_j;\hat{\boldsymbol{\psi}}^{(k)})\, f_{\text{SSMSN}}\left(y_j;\, \tilde{\boldsymbol{x}}_j^\top \hat{\boldsymbol{\varphi}}_l^{(k)},\, \hat{\sigma}_l^{2(k)},\, \hat{\lambda}_l^{(k)},\, \hat{\boldsymbol{\nu}}_l^{(k)}\right)},$$

$$\widehat{W}_{1ij}^{(k)} = \mathrm{E}(\tau_{1j} \mid y_j, \boldsymbol{\theta}_i^{(k)}), \quad \widehat{W}_{2ij}^{(k)} = \mathrm{E}(\tau_{2j} \mid y_j, \boldsymbol{\theta}_i^{(k)}), \quad \widehat{W}_{3ij}^{(k)} = \mathrm{E}(\tau_{2j}\gamma_j \mid y_j, \boldsymbol{\theta}_i^{(k)}).$$

In the Appendix, we present the exact formulas required to compute the conditional expectations for the specific instances of the MoER-SSMSN family utilized in this study. During the $(k+1)$-th iteration, the parameter estimates are updated by differentiating the function Q. The steps involved in executing the ECME algorithm are summarized as follows:

E-step: Consider $\boldsymbol{\Theta} = \boldsymbol{\Theta}^{(k)}$, obtain $\hat{z}_{ij}^{(k)}$, $\widehat{W}_{1ij}^{(k)}$, $\widehat{W}_{2ij}^{(k)}$, and $\widehat{W}_{3ij}^{(k)}$ for $j = 1, \ldots, n$ and $i = 1, \ldots, g$.

CM-step 1: Maximizing equation (3.6) with respect to $\boldsymbol{\varphi}_i$ yields the updated estimate $\hat{\boldsymbol{\varphi}}_j^{(k)}$

$$\hat{\boldsymbol{\varphi}}_i^{(k+1)} = \left(\sum_{j=1}^{n} \hat{z}_{ij}^{(k)} \left(\widehat{W}_{1ij}^{(k)} + \hat{\lambda}_i^{2(k)} \widehat{W}_{2ij}^{(k)}\right) \boldsymbol{x}_j \boldsymbol{x}_j^\top\right)^{-1} \left(\sum_{j=1}^{n} \hat{z}_{ij}^{(k)} \left(\left(\widehat{W}_{1ij}^{(k)} + \hat{\lambda}_i^{2(k)} \widehat{W}_{2ij}^{(k)}\right) y_j - \hat{\sigma}_i^{(k)} \hat{\lambda}_i^{(k)} \widehat{W}_{3ij}^{(k)}\right) \boldsymbol{x}_j\right).$$

CM-step 2: By fixing $\boldsymbol{\varphi}_i = \hat{\boldsymbol{\varphi}}_i^{(k+1)}$ and maximizing equation (3.6) with respect to σ_i^2, we obtain the updated estimate $\hat{\sigma}_i^{2(k)}$

$$\hat{\sigma}_i^{2(k+1)} = \frac{\sum_{j=1}^{n} \hat{z}_{ij}^{(k)} \widehat{W}_{1ij}^{(k)} \left(y_j - \hat{\boldsymbol{\varphi}}_i^{(k+1)^\top} \tilde{\boldsymbol{x}}_j\right)^2}{\sum_{j=1}^{n} \hat{z}_{ij}^{(k)}}.$$

CM-step 3: By fixing $\boldsymbol{\varphi}_i = \hat{\boldsymbol{\varphi}}_i^{(k+1)}$ and maximizing equation (3.6) with respect to η_i, we obtain the updated estimate $\hat{\eta}_j^{(k)}$ by

$$\hat{\eta}_i^{(k+1)} = \frac{\sum_{j=1}^{n} \hat{z}_{ij}^{(k)} \widehat{W}_{3ij}^{(k)} \left(y_j - \hat{\boldsymbol{\varphi}}_i^{(k+1)^\top} \tilde{\boldsymbol{x}}_j\right)}{\sum_{j=1}^{n} \hat{z}_{ij}^{(k)} \widehat{W}_{2ij}^{(k)} \left(y_j - \hat{\boldsymbol{\varphi}}_i^{(k+1)^\top} \tilde{\boldsymbol{x}}_j\right)^2},$$

To compute the updated value of $\hat{\lambda}^{(k+1)}$, we use the following formula:

$$\hat{\lambda}^{(k+1)} = \frac{\hat{\eta}_i^{(k+1)} \hat{\sigma}_i^{(k+1)} \sum_{j=1}^{n} \hat{z}_{ij}^{(k)} \widehat{W}_{3ij}^{(k)} \hat{u}_{ij}^{(k+1)}}{\sum_{j=1}^{n} \hat{z}_{ij}^{(k)} \widehat{W}_{2ij}^{(k)} \left(\hat{u}_{ij}^{(k+1)}\right)^2},$$

where $\hat{u}_j^{(k+1)} = \frac{y_j - \hat{\boldsymbol{\varphi}}_i^{(k+1)^\top} \tilde{\boldsymbol{x}}_j}{\hat{\sigma}^{(k+1)}}$.

CM-step 4: Based on Nguyen and McLachlan (2016), the update for $\hat{\boldsymbol{\psi}}_i^{(k)}$ is given by:

$$\hat{\boldsymbol{\psi}}_i^{(k+1)} = 4\left(\sum_{j=1}^{n} \boldsymbol{w}_j \boldsymbol{w}_j^\top\right)^{-1} \sum_{j=1}^{n} \boldsymbol{w}_j \left(\hat{z}_{ij}^{(k)} - \pi_i(\boldsymbol{w}_j; \hat{\boldsymbol{\psi}}^{(k)})\right) + \hat{\boldsymbol{\psi}}_j^{(k)}, \quad i = 1, \ldots, g.$$

For updating $\hat{\boldsymbol{\nu}}_i^{(k)}$, it is closely linked to the form of $h(\boldsymbol{\tau}_j; \boldsymbol{\nu}_i)$. Since evaluating

$$\mathrm{E}\left(\log\left(\sqrt{\tau_{1j}\tau_{2j}} h(\tau_{1j}, \tau_{2j}; \boldsymbol{\nu}_i)\right) \mid y_j, \boldsymbol{\theta}_i^{(k)}\right),$$

can be complex, an alternative is to maximize the restricted actual log-likelihood function, which leads regarding the subsequent CML-step:

$$\hat{\boldsymbol{\nu}}_i^{(k+1)} = \arg\max_{\boldsymbol{\nu}_i} \sum_{j=1}^{n} \hat{z}_{ij}^{(k)} \log f_{\mathrm{SSMSN}}\left(y_j; \hat{\boldsymbol{\varphi}}_i^{(k+1)^\top} \tilde{\boldsymbol{x}}_j, \hat{\sigma}_i^{2(k+1)}, \hat{\lambda}_i^{(k+1)}, \boldsymbol{\nu}_i\right). \tag{3.7}$$

3.3 ECME approach to estimating the STT Distribution

Based on Appendix A, for the STT distribution, we observe that

$$h(\boldsymbol{\tau}; \boldsymbol{\nu}) = f_\Gamma\left(\tau_1; \frac{\nu_1}{2}, \frac{\nu_1}{2}\right) f_\Gamma\left(\tau_2; \frac{\nu_2}{2}, \frac{\nu_2}{2}\right),$$

As stated in equation (3.6), the E-step involves the computation of two additional conditional expectations:

$$\widehat{W}_{4ij}^{(k)} = \mathrm{E}\left(\log \tau_{1j} \mid y_j, \hat{\boldsymbol{\theta}}_i^{(k)}\right), \qquad \widehat{W}_{5ij}^{(k)} = \mathrm{E}\left(\log \tau_{2j} \mid y_j, \hat{\boldsymbol{\theta}}_i^{(k)}\right),$$

which can be evaluated using Appendix A.

CM-step 5: The parameters ν_{1i} and ν_{2i} are updated by maximizing equation (3.6) with respect to them. This leads to solve for the roots of the following equations:

$$\log\left(\frac{\nu_{1i}}{2}\right) - \mathrm{DG}\left(\frac{\nu_{1i}}{2}\right) + 1 + \frac{1}{\sum_{i=1}^{n} \hat{z}_{ij}^{(k)}} \sum_{i=1}^{n} \hat{z}_{ij}^{(k)} \left(\widehat{W}_{4ij}^{(k)} - \widehat{W}_{1ij}^{(k)}\right) = 0,$$

$$\log\left(\frac{\nu_{2i}}{2}\right) - \mathrm{DG}\left(\frac{\nu_{2i}}{2}\right) + 1 + \frac{1}{\sum_{i=1}^{n} \hat{z}_{ij}^{(k)}} \sum_{i=1}^{n} \hat{z}_{ij}^{(k)} \left(\widehat{W}_{5ij}^{(k)} - \widehat{W}_{2ij}^{(k)}\right) = 0.$$

Alternatively, the updates for $\hat{\nu}_{1i}^{(k)}$ and $\hat{\nu}_{2i}^{(k)}$ can be obtained by regarding the subsequent CML-step:

CML-step:

$$(\hat{\nu}_{1i}^{(k+1)}, \hat{\nu}_{2i}^{(k+1)}) = \arg\max_{(\nu_1,\nu_2)} \sum_{j=1}^{n} \hat{z}_{ij}^{(k)} \log f_{\mathrm{STT}}\Big(y_j; \hat{\boldsymbol{\varphi}}_i^{(k+1)^\top} \tilde{\boldsymbol{x}}_j, \hat{\sigma}_i^{2(k+1)}, \hat{\lambda}_i^{(k+1)}, \nu_1, \nu_2\Big).$$

In many real-world scenarios, it is common to impose the condition $\nu_{21} = \nu_{22} = \cdots = \nu_{2g} = \nu_2$ to make sure the stability and convergence of the estimation algorithm. Following a procedure similar to that in equation (3.7), the update of $\widehat{\nu}_2^{(k)}$ is obtained through the following optimization step:

$$\hat{\nu}_2^{(k+1)} = \arg\max_{\nu_2} \left\{ \sum_{i=1}^{g} \sum_{j=1}^{n} \widehat{z}_{ij}^{(k+1)} \log f_{\mathrm{STT}}\left(y_j; \hat{\alpha}_i^{(k+1)}, \hat{\boldsymbol{\varphi}}_i^{(k+1)^\top} \tilde{\boldsymbol{x}}, \hat{\lambda}_i^{(k+1)}, \hat{\nu}_{1i}^{(k+1)}, \nu_2\right) \right\}.$$

3.4 ECME approach to estimating the ST Distribution

For the ST distribution, we recall from equation (A.9) that

$$h(\tau; \nu) = f_{\Gamma}\left(\tau; \frac{\nu}{2}, \frac{\nu}{2}\right),$$

Accordingly, the E-step in equation (3.6) simplifies to

$$\widehat{W}_{1ij}^{(k)} = \widehat{W}_{2ij}^{(k)} = \widehat{W}_{ij}^{(k)} = \mathrm{E}\left(\tau_j \mid y_j, \hat{\boldsymbol{\theta}}_i^{(k)}\right),$$

and includes an along with conditional expectation,which is expressed as

$$\widehat{W}_{4ij}^{(k)} = \mathrm{E}\left(\log \tau_j \mid y_j, \hat{\boldsymbol{\theta}}_i^{(k)}\right)$$

that can be computed by applying the formula given in Appendix A.

CM-step 5: To update $\nu_i^{(k)}$, we must find the solution to the following equation:

$$\log\left(\frac{\nu_i}{2}\right) - DG\left(\frac{\nu_i}{2}\right) + 1 + \frac{1}{\sum_{i=1}^{n} \hat{z}_{ij}^{(k)}} \sum_{i=1}^{n} \hat{z}_{ij}^{(k)} \left(\widehat{W}_{4ij}^{(k)} - \widehat{W}_{ij}^{(k)}\right) = 0,$$

An alternative approach to determine $\hat{\nu}_i^{(k)}$ is by maximizing the constrained log-likelihood function, which is achieved by executing the following CML-step:

$$\hat{\nu}_i^{(k+1)} = \arg\max_{\nu_1} \sum_{j=1}^{n} \hat{z}_{ij}^{(k)} \log f_{\mathrm{ST}}\Big(y_j; \hat{\boldsymbol{\varphi}}_i^{(k+1)^\top} \tilde{\boldsymbol{x}}_j, \hat{\sigma}_i^{2(k+1)}, \hat{\lambda}_i^{(k+1)}, \nu\Big).$$

In many real-world scenarios, it is common to impose the condition $\nu_1 = \nu_2 = \cdots = \nu_g = \nu$ to ensure the stability and convergence of the estimation algorithm. Following

a procedure similar to that in equation (3.7), the update of $\hat{\nu}_2^{(k)}$ is obtained through the following optimization step:

$$\hat{\nu}_2^{(k+1)} = \arg\max_{\nu_2} \left\{ \sum_{i=1}^{g} \sum_{j=1}^{n} \hat{z}_{ij}^{(k+1)} \log f_{\mathrm{ST}}\left(y_j; \hat{\alpha}_i^{(k+1)}, \hat{\boldsymbol{\varphi}}_i^{(k+1)^\top} \tilde{\boldsymbol{x}}, \hat{\lambda}_i^{(k+1)}, \nu\right) \right\}.$$

3.5 ECME approach to estimating the SGLN Distribution

In A, it is explained that the SGLN distribution arises from the SSMSN model by setting τ_{1j} as $\tau_{1j} \overset{d}{=} \frac{1}{\chi^2_{(2\upsilon)}}$, while τ_{2j} equals 1 with probability 1, for each $j = 1, \ldots, n$. Accordingly, we obtain

$$h(\boldsymbol{\tau}_j; \boldsymbol{\nu}) = h(\tau_j; \boldsymbol{\nu}) = \frac{1}{2^{\upsilon}\Gamma(\upsilon)} \tau_j^{-(\upsilon+1)} \exp\left(-\frac{1}{2\tau_j}\right).$$

Then, the E-step in equation (3.6) becomes

$$\widehat{W}_{1ij}^{(k)} = \mathrm{E}(\tau_j \mid y_j, \hat{\boldsymbol{\theta}}_i^{(k)}), \quad \widehat{W}_{2ij}^{(k)} = 1, \quad \widehat{W}_{3ij}^{(k)} = \mathrm{E}(\gamma_j \mid y_j, \hat{\boldsymbol{\theta}}_i^{(k)}).$$

and it also requires one additional conditional expectation given by

$$\widehat{W}_{4ij}^{(k)} = \mathrm{E}\left(\log \tau_j \mid y_j, \hat{\boldsymbol{\theta}}_i^{(k)}\right)$$

which can be computed using Appendix A.

CM-step 5: The parameter $\nu_i^{(k)}$ is updated by maximizing equation (3.6) with respect to them. This leads to solving for the roots of the following equations:

$$\log(2) + DG(\nu_i) + \frac{1}{\sum_{j=1}^{n} \hat{z}_{ij}^{(k)}} \sum_{j=1}^{n} \hat{z}_{ij}^{(k)} \widehat{W}_{4ij}^{(k)} = 0.$$

An alternative approach for updating $\hat{\nu}_i^{(k)}$ without the need to compute $\widehat{W}_{4ij}^{(k)}$ is given by

CML-step:

$$\hat{\nu}_i^{(k+1)} = \arg\max_{\nu_i} \left\{ \sum_{j=1}^{n} \hat{z}_{ij}^{(k+1)} \log f_{\mathrm{SGLN}}\left(y_j; \hat{\alpha}_i^{(k+1)}, \hat{\boldsymbol{\varphi}}_i^{(k+1)^\top} \tilde{\boldsymbol{x}}, \hat{\lambda}_i^{(k+1)}, \nu_i\right) \right\}.$$

3.6 ECME approach to estimating the SSN Distribution

As explained in Appendix A, the SSN distribution is derived as a particular case of the SSMSN family by setting $\tau_{1j} = \tau_j \sim \mathrm{Beta}\left(\frac{\nu_i}{2}, 1\right)$ and $\tau_{2j} = 1$ with probability one, for $i = 1, \ldots, n$. Hence, the pdf of SSN can be obtain by

$$h(\tau_j; \nu_i) = \nu_i \, \tau_j^{\nu_i - 1}.$$

Accordingly, the E-step in (3.6) reduces to

$$\widehat{W}_{1ij}^{(k)} = \mathrm{E}(\tau_j \mid y_j, \hat{\boldsymbol{\theta}}_i^{(k)}),\ \widehat{W}_{2ij}^{(k)} = 1,\ \widehat{W}_{3ij}^{(k)} = \mathrm{E}(\gamma_j \mid y_j, \hat{\boldsymbol{\theta}}_i^{(k)}),\ \widehat{W}_{4ij}^{(k)} = \mathrm{E}\left(\log \tau_j \mid y_j, \hat{\boldsymbol{\theta}}_i^{(k)}\right)$$

where $\widehat{W}_{4ij}^{(k)}$ can be computed using Appendix A. Finally, the **CM-step 5** updates $\nu_i^{(k)}$ according to

$$\nu_i^{(k+1)} = -\frac{2\sum_{j=1}^{n} \hat{z}_{ij}^{(k)}}{\sum_{j=1}^{n} \hat{z}_{ij}^{(k)} \widehat{W}_{4ij}^{(k)}}.$$

An alternative approach for updating $\hat{\nu}_i^{(k)}$ without the need to compute $\widehat{W}_{4ij}^{(k)}$ is given by

CML-step:

$$\hat{\nu}_i^{(k+1)} = \arg\max_{\nu_i} \left\{ \sum_{j=1}^{n} \hat{z}_{ij}^{(k+1)} \log f_{\mathrm{SSN}}\left(y_j; \hat{\alpha}_i^{(k+1)}, \hat{\boldsymbol{\varphi}}_i^{(k+1)^\top} \tilde{\boldsymbol{x}}, \hat{\lambda}_i^{(k+1)}, \nu_i\right) \right\}.$$

3.7 Standard errors via observed information

The information matrix in statistical models is a key tool for assessing the accuracy of parameter estimates. It is typically defined as the negative second derivative of the log-likelihood function with respect to the model parameters. The information matrix provides insights into how the likelihood function changes around the estimated parameter values and is particularly useful for determining the standard errors of the estimates and evaluating the model's fit. The standard errors calculated from this matrix are crucial for conducting significance tests on parameters and constructing their corresponding confidence intervals. Furthermore, these errors help in pinpointing the variables that significantly influence the classification. However, deriving the second derivatives analytically for the MoER-SSMSN model proves to be quite intricate and computationally demanding. To overcome this challenge, we follow the approach in Basford et al. (1997), where the outer product of the individual score (gradient) vectors is used as an efficient approximation for the observed information matrix, which can be written as

$$\hat{I}_e = \sum_{j=1}^{n} \hat{s}_j \hat{s}_j^\top, \tag{3.8}$$

where

$$\hat{s}_j = \frac{\partial f(y_j; \boldsymbol{\Theta})}{\partial \boldsymbol{\Theta}} = \mathrm{E}_{\widehat{\boldsymbol{\Theta}}}\left(\frac{\partial \ell_{cj}(\boldsymbol{\Theta} \mid y_j, \gamma_j, \tau_{1j}, \tau_{2j}, Z_j)}{\partial \boldsymbol{\Theta}} \middle| y_j \right)$$

The individual score vector for each observation, indexed by $j = 1, \ldots, n$, is derived based on the findings of Louis (1982). In this context, $\ell_{cj}(\boldsymbol{\Theta} \mid y_j, \gamma_j, \tau_{1j}, \tau_{2j}, Z_j)$ represents the log-likelihood function of the complete data for the jth observation. Therefore, the following expression holds:

$$\ell_{cj}(\boldsymbol{\Theta} \mid y_j, \gamma_j, \tau_{1j}, \tau_{2j}, Z_j) = \sum_{i=1}^{g} Z_{ij}\Big[\log \pi_i(\boldsymbol{w}_j; \boldsymbol{\psi}) - \frac{\tau_{1j}(y_j - \boldsymbol{\varphi}_i^\top \tilde{\boldsymbol{x}}_j)^2}{2\sigma_i^2} + \eta_i \tau_{2j} \gamma_j (y_j - \boldsymbol{\varphi}_i^\top \tilde{\boldsymbol{x}}_j) - \frac{\eta_i^2 \tau_{2j} (y_j - \boldsymbol{\varphi}_i^\top \tilde{\boldsymbol{x}}_j)^2}{2} - \log(\sigma_i) + \log\Big(\sqrt{\tau_{1j}\tau_{2j}} h(\tau_{1j}, \tau_{2j}; \boldsymbol{\nu}_i)\Big)\Big]. \tag{3.9}$$

For further clarification, the gradient vector $\hat{s}_j$ is composed of the following partial derivatives:

$$\hat{s}_j = \Big(\hat{s}_{j,\boldsymbol{\psi}_1}, \ldots, \hat{s}_{j,\boldsymbol{\psi}_{g-1}}, \hat{s}_{j,\alpha_1}, \ldots, \hat{s}_{j,\alpha_g}, \hat{s}_{j,\boldsymbol{\varphi}_1}, \ldots, \hat{s}_{j,\boldsymbol{\varphi}_g}, \hat{s}_{j,\lambda_1}, \ldots, \hat{s}_{j,\lambda_g}, \hat{s}_{j,\boldsymbol{\nu}_1}, \ldots, \hat{s}_{j,\boldsymbol{\nu}_g}\Big).$$

By applying vector and matrix differentiation to (3.9), we derive:

$$\begin{aligned}
\hat{s}_{j,\boldsymbol{\psi}_i} &= (\hat{z}_{ij} - \pi_i(\boldsymbol{w}_j; \hat{\boldsymbol{\psi}}))\boldsymbol{w}_j, \\
\hat{s}_{j,\alpha_i} &= \hat{z}_{ij}\left[-\frac{\hat{\lambda}_i \widehat{W}_{3ij} \zeta_2(y_j, \hat{\boldsymbol{\varphi}}_i)}{\hat{\alpha}_i^2} + \frac{\hat{\lambda}_i^2 \widehat{W}_{2ij} \zeta_2^2(y_j, \hat{\boldsymbol{\varphi}}_i) - \widehat{W}_{1ij} \zeta_2^2(y_j, \hat{\boldsymbol{\varphi}}_i)}{\hat{\alpha}_i^3} - \frac{1}{\hat{\alpha}_i}\right], \\
\hat{s}_{j,\boldsymbol{\varphi}_i} &= \hat{z}_{ij}\left[\frac{\hat{\lambda}_i \widehat{W}_{3ij} \frac{\partial \zeta_2(y_j, \hat{\boldsymbol{\varphi}}_i)}{\partial \boldsymbol{\varphi}_i}}{\hat{\alpha}_i} - \frac{\zeta_2(y_j, \hat{\boldsymbol{\varphi}}_i) \frac{\partial \zeta_2(y_j, \hat{\boldsymbol{\varphi}}_i)}{\partial \boldsymbol{\varphi}_i} (\hat{\lambda}_i^2 \widehat{W}_{2ij} - \widehat{W}_{1ij})}{\hat{\alpha}_i^2} + \frac{\frac{\partial \zeta_2(y_j, \hat{\boldsymbol{\varphi}}_i)}{\partial \boldsymbol{\varphi}_i}}{\zeta_2(y_j, \hat{\boldsymbol{\varphi}}_i)}\right], \\
\hat{s}_{j,\lambda_i} &= \hat{z}_{ij}\left[\frac{\widehat{W}_{3ij} \zeta_2(y_j, \hat{\boldsymbol{\varphi}}_i)}{\hat{\alpha}_i} - \frac{\hat{\lambda}_i \widehat{W}_{2ij} \zeta_2^2(y_j, \hat{\boldsymbol{\varphi}}_i)}{\hat{\alpha}_i^2}\right].
\end{aligned} \tag{3.10}$$

It should be noted that conditional expectation estimates, including $\hat{z}_{ij}$, are defined similarly to previous ones, with $\hat{\boldsymbol{\Theta}}$ replacing $\hat{\boldsymbol{\Theta}}^{(k)}$. However, it should be noted that the forms of these conditional expectations vary according to the specific subfamily of the SSMSN distributions considered, as described in the corresponding subsections.

The detailed expressions for the components of $\hat{s}_{j,\boldsymbol{\nu}_i}$ are provided in the following.

(i) STT model:

$$\begin{aligned}
\hat{s}_{j,\nu_{1i}} &= \frac{\hat{z}_{ij}}{2}\left[\log\left(\frac{\hat{\nu}_{1i}}{2}\right) + 1 + \widehat{W}_{4ij} - \widehat{W}_{1ij} - DG\left(\frac{\hat{\nu}_{1i}}{2}\right)\right], \\
\hat{s}_{j,\nu_{2i}} &= \frac{\hat{z}_{ij}}{2}\left[\log\left(\frac{\hat{\nu}_{2i}}{2}\right) + 1 + \widehat{W}_{5ij} - \widehat{W}_{2ij} - DG\left(\frac{\hat{\nu}_{2i}}{2}\right)\right].
\end{aligned}$$

If $\nu_{21} = \cdots = \nu_{2g} = \nu_2$, then $\hat{s}_{j,\nu_2} = \sum_{i=1}^{g} \hat{s}_{j,\nu_{2i}}$.

(ii) SGLN model:

$$\hat{s}_{j,\nu_i} = -\hat{z}_{ij}\left[\log 2 + DG(\hat{\nu}_i) + \widehat{W}_{4ij}\right].$$

(iii) SSN model:

$$\hat{s}_{j,\nu_i} = \hat{z}_{ij}\left(\frac{1}{\hat{\nu}_i} + \frac{\widehat{W}_{4ij}}{2}\right).$$

Assuming regularity conditions are met, the standard errors for the estimates of $\hat{\boldsymbol{\Theta}}$ are obtained by taking the square roots of the diagonal elements from the inverse of the information matrix, as described in equation (3.8).

3.8 Initialization Procedure and Convergence Assessment

The EM-type algorithms inherently suffer from the limitation that they do not guarantee convergence to the global maximum, owing to the multimodal structure of the likelihood surface in mixture-based models. In fact, their efficiency is largely influenced by the choice of initial parameter values (Baudry and Celeux, 2015). To mitigate this issue and to improve numerical stability, we recommend the following initialization strategy for the proposed ECME algorithm:

1. Combine the response vector and the corresponding covariates from the expert networks into a single augmented matrix. The matrix is then partitioned into g clusters using the k-means algorithm implemented in the `KMeans_rcpp()` function in R, initialized through the k-means++ strategy.

2. For initializing the gating coefficients $\boldsymbol{\psi}_i$, two simple schemes may be employed. In the first approach, set $\boldsymbol{\psi}_i = \mathbf{0}_p$ for $i = 1, \ldots, g-1$, which effectively reduces the MoE model to equal mixing proportions. Alternatively, one may compute the initial posterior probabilities $\hat{z}_{ij}^{(0)}$ from the k-means classification results and estimate $\boldsymbol{\psi}$ by fitting a generalized linear model.

3. Using the cluster assignments obtained from k-means, estimate the regression parameters $\hat{\boldsymbol{\varphi}}_i^{(0)}$ by fitting separate linear regression models within each cluster based on Santana et al. (2011).

4. Two possible choices may be adopted for initializing the skewness parameters. One can either assign $\hat{\lambda}_i^{(0)} = 0$ for all i, or initialize them using the empirical

skewness of the residuals in each cluster. It is advisable to implement both initialization schemes and retain the model corresponding to the higher log-likelihood value.

5. The scale parameters $\hat{\sigma}_i^{2(0)}$ are obtained by

$$\hat{\sigma}_i^{2(0)} = \frac{\sum_{j=1}^{n} \hat{z}_{ij}^{(0)} (y_j - \hat{\boldsymbol{\varphi}}_i^{(0)^\top} \tilde{\boldsymbol{x}}_j)^2}{\sum_{j=1}^{n} \hat{z}_{ij}^{(0)}}.$$

6. Specify relatively small initial values for $\hat{\boldsymbol{\nu}}_i^{(0)}$.

The EM-type algorithm is monitored to convergence through the relative change in the observed log-likelihood, defined as

$$\frac{\ell(\hat{\boldsymbol{\Theta}}^{(k+1)}) - \ell(\hat{\boldsymbol{\Theta}}^{(k)})}{\ell(\hat{\boldsymbol{\Theta}}^{(k)})} < \varepsilon,$$

where ε represents a user-specified tolerance level. During our experiments, convergence was considered achieved either when the maximum number of iterations ($k_{\max} = 2000$) was reached or when the relative change in the log-likelihood dropped below 10^{-5}.

The proposed estimation procedure follows the standard ECME framework. Therefore, under regularity conditions, the observed log-likelihood is guaranteed to be non-decreasing across iterations. The convergence properties of EM-type algorithms have been extensively studied in the literature; see, for example, Dempster et al. (1977) and Liu and Rubin (1994). Furthermore, the convergence behavior of the proposed EM-type algorithm was examined under several different initial values of the parameters. The numerical results demonstrated stable convergence patterns with nearly identical final log-likelihood values, indicating low sensitivity of the algorithm to parameter initialization.

4 Risk Evaluation for MoER-SSMSN Models

Evaluating risk is of paramount importance for investors managing portfolios containing risky assets. Consequently, risk metrics and their underlying theories are critical for estimating potential financial losses. In practice, these risk measures serve several

key functions, such as determining the required risk capital, ensuring that capital requirements are met, providing tools for effective management, and setting insurance premiums McNeil et al. (2015). Statistical methods are essential in achieving these goals, as the majority of modern risk measures applied to portfolios are inherently statistical. The following proposition outlines how various risk measures can be derived, assuming that asset characteristics adhere to the SSMSN distribution.

In this context, several risk metrics are considered, including the Target Shortfall (TS), Value-at-Risk (VaR), and, notably, the Tail-Value-at-Risk (TVaR). For clarity and simplicity in our notation, we define the following: the r-th order upper partial moment of a random variable, along with the probabilities of shortfall (PS) and outperformance (PO).

Let $E[Y-y_q]_+^r$ represent the upper partial moment of order r for the random variable Y, with respect to $\mathbb{R}^+$. In particular, this can be written as

$$E[Y-y_q]_+^r = \int_{t_q}^{\infty} (y-t_q)^r f_Y(y;\omega)\, dy,$$

where y_q acts as a reference point that distinguishes between accretions and depletions, and $f_Y(\cdot;\omega)$ denotes the pdf of Y, parameterized by ω. The reference point y_q can either be a static threshold, such as a predetermined income poverty line applicable to all households, or a flexible target that changes based on the distribution of the random variable specific to each household. In the next proposition, we derive explicit formulas for calculating the probabilities of shortfall (PS) and outperformance (PO) within the SSMSN framework.

Proposition 4.1. *Consider the random variable Y following the distribution $Y \sim SSMSN(\mu, \sigma^2, \lambda, \boldsymbol{\nu})$. The probabilities of Y either falling below or exceeding a given threshold y_q are expressed as follows: The probability of Y falling below the target y_q, known as the probability of shortfall, is given by*

$$PS(y_q, \mu, \sigma^2, \lambda, \boldsymbol{\nu}) = 1 - E[Y-y_q]_{0+} = F_{SSMSN}(y_q; \mu, \sigma^2, \lambda, \boldsymbol{\nu}),$$

while the probability of outperformance, where Y exceeds the target, is

$$PO(y_q, \mu, \sigma^2, \lambda, \boldsymbol{\nu}) = E[Y-y_q]_{0+} = 1 - F_{SSMSN}(y_q; \mu, \sigma^2, \lambda, \boldsymbol{\nu}).$$

Here, $F_{SSMSN}(\cdot; \mu, \sigma^2, \lambda, \boldsymbol{\nu})$ represents the cdf of the SSMSN distribution, standardized appropriately.

Proof. The proof is straightforward and therefore not included. □

It is important to note that closed-form phrases for PS and PO of the SSMSN distribution do not exist. As a result, risk evaluations are typically conducted using the cdf of the SSMSN distribution. The target shortfall (TS) risk measure, which quantifies the deviation of a random variable from a specified threshold, is characterized as the first-order upper partial moment with respect to $y_q \in \mathbb{R}^+$.

For a random variable $Y \sim \text{SSMSN}(\mu, \sigma^2, \lambda, \boldsymbol{\nu})$, the TS measure can be expressed as follows:

$$TS(y_q, \mu, \sigma^2, \lambda, \boldsymbol{\nu}) = E[Y - y_q]_+ = \int_{y_q}^{\infty} (y - y_q) f_{\text{SSMSN}}(y; \mu, \sigma^2, \lambda, \boldsymbol{\nu})\, dy.$$

Alternatively, this can be reformulated as:

$$TS(y_q, \mu, \sigma^2, \lambda, \boldsymbol{\nu}) = \int_{y_q}^{\infty} y f_{\text{SSMSN}}(y; \mu, \sigma^2, \lambda, \boldsymbol{\nu})\, dy - y_q\, PO(y_q, \mu, \sigma^2, \lambda, \boldsymbol{\nu}).$$

The Value-at-Risk (VaR) is a commonly used measure of risk in financial markets. For a given confidence level $q \in (0, 1)$, VaR is defined as the value below which the probability of loss is at most q:

$$\text{VaR}_q(Y) = \sup \{y \mid F_Y(y) \leq q\},$$

where $F_Y(y)$ is the cdf of Y. However, VaR has been questioned for its loss of consistency, as is heavily influenced by the tail of the loss distribution, which can result in misleading assessments of risk.

In contrast, the Tail-Value-at-Risk (TVaR) is considered a more reliable and coherent risk measure. It satisfies important properties such as regularity, additive property, sameness, and transferability across shifts, and represents the anticipated extreme loss. TVaR delivers the foreseen value of losses that exceed a considering threshold. For every $q \in (0, 1)$, the definition of TVaR is:

$$\text{TVaR}(Y; q) = E[Y \mid Y \geq \text{VaR}_q(Y)].$$

The TVaR can also be characterized by the tail conditional expectation, as follows:

$$\text{TVaR}(T; q) = \frac{1}{1 - q} \int_{y_q}^{\infty} y f_{\text{SSMSN}}(y; \mu, \sigma^2, \lambda, \boldsymbol{\nu})\, dy,$$

where y_q is relative to the distribution with q-th percentile.

Corollary 4.2. *If Y follows a mixture of SSMSN distributions with the probability density function given by (3.2), and if $VaR_q(Y) = y_q$, therefore, the TVaR of Y is calculated as:*

$$TVaR(Y;q) = \frac{1}{1-q}\sum_{i=1}^{g}\pi_i(\boldsymbol{w};\boldsymbol{\psi})\int_{y_q}^{\infty} y f_{\text{SSMSN}}\left(y;\boldsymbol{\varphi}_i^{\top}\tilde{\boldsymbol{x}},\sigma_i^2,\lambda_i,\boldsymbol{\nu}_i\right).$$

5 Simulation studies

5.1 Finite-sample performance

In this part of the study, we perform a Monte Carlo (MC) simulation to explore the behavior of the proposed estimators derived from the ECME algorithm when applied to small sample sizes. Additionally, we assess the accuracy of the standard error estimates, which are calculated using the information-based approach described in Section 3.7. We generate 500 MC samples for each of the sample sizes $n = 50, 100, 250, 500, 1000$, and 2000, using the three-component MoER-STT models. The remaining model parameters are set as follows: $\boldsymbol{\psi}_1 = (2,-4)^{\top}$, $\boldsymbol{\psi}_2 = (-2,-4)^{\top}$, $\lambda_1 = 2$, $\lambda_2 = 3$, $\lambda_3 = -2$, $(\sigma_1^2,\sigma_2^2,\sigma_3^2) = (0.5,1,1.5)$, $(\nu_1,\nu_2,\nu_3) = (4,7,10)$, $\boldsymbol{\varphi}_1 = (0,4)^{\top}$, $\boldsymbol{\varphi}_2 = (0,-3)^{\top}$, and $\boldsymbol{\varphi}_3 = (-2,1)^{\top}$. The covariates are defined as $\boldsymbol{v}_i = \boldsymbol{x}_i = (1, x_{1i})^{\top}$, where x_{1i} is drawn from a uniform distribution $U(-2, 2)$.

In order to assess the precision of the estimates, we compute the absolute relative bias (ARB) and root mean square error (RMSE) through numerous replications, using the following formulas:

$$\text{ARB} = \frac{1}{500}\sum_{r=1}^{500}\left|\frac{\hat{\theta}_r - \theta_{\text{true}}}{\theta_{\text{true}}}\right|, \qquad \text{RMSE} = \sqrt{\frac{1}{500}\sum_{r=1}^{500}\left(\hat{\theta}_r - \theta_{\text{true}}\right)^2},$$

where $\hat{\theta}_r$ denotes the estimated parameter in the rth replication, and θ_{true} is its true value.

The first simulation study was designed to investigate the finite-sample behavior of the proposed estimators under different sample sizes. In particular, the empirical standard deviations (STD) and the average estimated standard errors (ASE) were evaluated in order to assess the stability and accuracy of the estimation procedure.

Additionally, to investigate the finite-sample behavior and stability of the proposed estimators, we compute the empirical standard deviations (STD) of the parameter estimates together with the average estimated standard errors (ASE) obtained from the

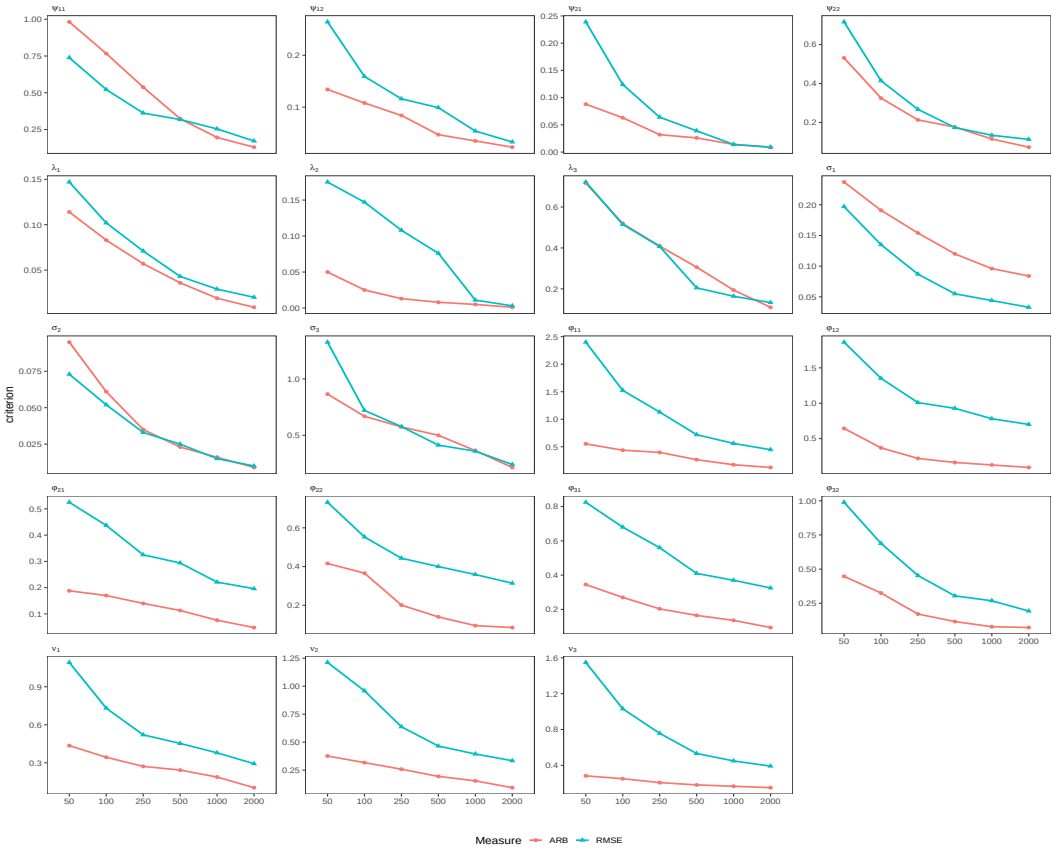

Figure 1: Absolute relative bias (ARB) and root mean squared error (RMSE) of parameter estimates across different sample sizes.

observed information matrix:

$$\text{STD} = \sqrt{\frac{1}{499}\sum_{r=1}^{500}\left(\hat{\theta}_r - \frac{1}{500}\sum_{s=1}^{500}\hat{\theta}_s\right)^2}, \qquad \text{ASE} = \frac{1}{500}\sum_{r=1}^{500} se(\hat{\theta}_r).$$

By comparing the values of STD and ASE, we can assess the quality of the asymptotic approximation in relation to the finite-sample variability of the estimators.

The simulation outcomes, illustrated in Figure 1, reveal a clear downward trend in both ARB and RMSE values with an increasing sample size n, strong empirical evidence is provided for the consistency of the proposed estimators. Moreover, the numerical evidence summarized in Table 1 shows that, as the sample size increases, the empirical standard deviations (STD) become progressively closer to the corresponding average

Table 1: Simulation outcomes for assessing the empirical standard deviations (STD) and the estimated standard errors (ASE) of parameter estimates for varying sample sizes.

n	Measure	ψ_{11}	ψ_{12}	ψ_{21}	ψ_{22}	λ_1	λ_2	λ_3	σ_1	σ_2	σ_3	φ_{11}	φ_{12}	φ_{21}	φ_{22}	φ_{31}	φ_{32}	ν_1	ν_2	ν_3
50	STD	1.060	0.305	0.295	0.370	0.960	0.215	0.210	0.305	1.550	2.300	0.675	0.755	0.625	0.705	0.080	0.380	0.350	0.330	0.300
	ASE	1.580	0.420	0.430	0.225	0.760	0.150	0.152	0.180	2.100	1.650	0.340	0.550	0.715	0.800	0.085	0.940	0.900	0.860	0.810
100	STD	0.875	0.245	0.240	0.295	0.790	0.175	0.180	0.260	1.280	1.950	0.570	0.645	0.515	0.585	0.070	0.280	0.255	0.235	0.215
	ASE	1.400	0.375	0.380	0.195	0.670	0.130	0.133	0.156	1.840	1.460	0.285	0.480	0.580	0.660	0.070	0.725	0.685	0.655	0.615
250	STD	0.510	0.125	0.127	0.155	0.415	0.095	0.097	0.120	0.830	1.140	0.335	0.350	0.275	0.310	0.045	0.210	0.190	0.170	0.155
	ASE	0.720	0.150	0.158	0.100	0.340	0.065	0.067	0.078	1.010	0.850	0.270	0.280	0.340	0.355	0.048	0.530	0.490	0.460	0.435
500	STD	0.300	0.065	0.063	0.080	0.250	0.055	0.055	0.070	0.480	0.675	0.240	0.245	0.200	0.215	0.035	0.145	0.135	0.120	0.105
	ASE	0.440	0.090	0.100	0.060	0.210	0.040	0.045	0.050	0.640	0.510	0.195	0.190	0.230	0.240	0.037	0.400	0.355	0.325	0.295
1000	STD	0.215	0.045	0.045	0.065	0.180	0.035	0.036	0.050	0.310	0.440	0.160	0.170	0.130	0.145	0.025	0.105	0.095	0.085	0.075
	ASE	0.280	0.045	0.046	0.060	0.150	0.035	0.036	0.045	0.335	0.390	0.150	0.155	0.150	0.160	0.025	0.270	0.240	0.220	0.195
2000	STD	0.140	0.030	0.028	0.037	0.115	0.027	0.027	0.034	0.215	0.325	0.105	0.115	0.085	0.098	0.018	0.075	0.068	0.060	0.054
	ASE	0.135	0.032	0.029	0.037	0.105	0.026	0.026	0.033	0.230	0.305	0.100	0.105	0.095	0.102	0.018	0.195	0.175	0.155	0.135

standard errors (ASE), confirming the adequacy of the approximation presented in Section 3.7 and indicating satisfactory finite-sample performance and consistency of the proposed estimators. Additional Monte Carlo experiments performed under alternative MoER-STT specifications exhibited comparable behavior, and are thus not reported here for the sake of conciseness. Taken together, these findings highlight the numerical stability, robustness, and overall reliability of the EM-based estimation framework across a variety of configurations.

5.2 Evaluating the effectiveness of models for heavy-tailed data

In this study, we assess the ability of the proposed model to handle data with heavy tails by evaluating its clustering performance. The experiment involved generating $n = 512$ observations using a three-component MoER-SSMSN model, based on the convolution representation of the SSMSN distribution. Two distinct schemes were used for generating the error terms ε_{ij} in equation (3.1). The first scenario (S1) assumed that the mixing variable ε_{ij} followed a mean–variance mixture of a skew-normal distribution, a model referred to as MoER-MVMSN. This model was not detailed in Section 2. In this scenario, the error terms were defined as $\varepsilon_{ij} = W_j\lambda_i + \sqrt{W_j}Z_{1j}$, where $Z_{1j} \sim \text{SN}(0, 1, 1.5)$ and $W \sim \text{GIG}(-\nu_i/2, \nu_i, 0)$, both drawn independently. The second scenario (S2) assumed that the expert components followed an ST distribution. To groups 1, 2, and 3, $x_{1i} \sim U(-3, -1)$, $x_{1i} \sim U(-2, 1)$, and $x_{1i} \sim U(0, 3)$, where $U(a, b)$ is uniform distribution with lower band a, upper band b. For the parameterization of the covariates, we set $\boldsymbol{w}_i = (1, w_{i1}, w_{i2})^\top$, where $w_{i1} \sim U(-2, 1)$ and $v_{i2} \sim N(0, 1.5)$. The gating parameters were

specified as $\boldsymbol{\psi}_1 = (2, -3, 5)^\top$ and $\boldsymbol{\psi}_2 = (1, -2.5, 9)^\top$. Additionally, the shape parameters were set to $\sigma_1 = 0.5$, $\sigma_2 = 1$, and $\sigma_3 = 2$, with skewness parameters $\lambda_1 = 3$, $\lambda_2 = -2$, and $\lambda_3 = -3$, and the degrees of freedom set to $\nu_1 = 2$, $\nu_2 = 3$, and $\nu_3 = 5$. The regression coefficient matrices were defined as $\boldsymbol{\varphi}_1 = (-4, 4)^\top$, $\boldsymbol{\varphi}_2 = (0, -3)^\top$, and $\boldsymbol{\varphi}_3 = (0, 4)^\top$.

We ran 100 simulation replications, each considering six submodels from the MoER-SSMSN family and their standard mixture counterparts (SSMSN models) for $g \in \{1, 2, 3, 4, 5\}$, leading to 60 fitted models per replication. The optimal number of expert components was determined using three well-known information criteria: the BIC (Schwarz, 1978), the ICL criterion (Biernacki et al., 2000), and the EDC (Bai et al., 1989). To evaluate the classification performance of the fitted models, we employed the ARI (Hubert and Arabie, 1985), the Jac. (Tanimoto, 1958), and the VI distance (Meilă, 2007). A model with a smaller VI value (closer to zero) implies higher clustering precision, whereas larger ARI and Jac. values (approaching one) reflect a stronger correspondence between the predicted clusters and the true class memberships. For further comparison, the proposed models were also evaluated against two related Mixture of Experts regression models introduced by Sepahdar et al. (2022), namely the MoER based on mean mixture of normal exponential distribution (MoER-MMNE) and based on mean mixture of normal exponential and half-normal distribution (MoER-MMNEH).

As reported in Table 2, the MoER-based models consistently outperform their standard counterparts, demonstrating the benefit of incorporating covariates into the mixing proportions. The MoER-SSMSN models show superior performance compared to both MoER-N and MR-N models. Moreover, models with heavy-tailed distributions such as STT and SGLN typically perform better across all three clustering metrics.

In Scenario 1 (S1), the MoER-STT model performs best in clustering, achieving the highest quality measure values. In Scenario 2 (S2), the MoER-STT and MoER-SGLN models show similar clustering accuracy, with slightly lower information criteria values for the MoER-SGLN model, suggesting a better fit. Furthermore, the three-component model yielded the best results for STT and SGLN distributions, whereas four or five-component models were more effective for the remaining distributions in terms of clustering performance.

Table 2: Average values of BIC, ICL, EDC, ARI, Jaccard coefficient, and VI across 100 Monte Carlo replications for evaluating the clustering performance of the proposed models under heavy-tailed scenarios.

	Scenario S1						Scenario S2					
Model	BIC	ICL	EDC	ARI	Jac.	VI	BIC	ICL	EDC	ARI	Jac.	VI
MoER-STT	-532.215	-536.592	-525.786	0.921	0.880	0.060	-490.032	-495.330	-480.246	0.951	0.930	0.050
MoER-SGLN	-511.146	-515.345	-505.472	0.912	0.870	0.070	-475.132	-480.610	-465.157	0.941	0.910	0.060
MoER-ST	-503.218	-505.447	-495.275	0.903	0.860	0.080	-460.145	-465.379	-460.139	0.932	0.900	0.065
MoER-SSN	-488.186	-485.336	-475.204	0.883	0.850	0.085	-455.108	-460.209	-455.183	0.925	0.890	0.070
MoER-MMNEH	-497.284	-500.416	-488.233	0.891	0.845	0.082	-458.327	-462.518	-450.286	0.921	0.882	0.068
MoER-MMNE	-491.172	-494.285	-482.164	0.887	0.838	0.084	-452.491	-456.672	-445.318	0.917	0.875	0.071
MoER-SN	-475.160	-475.287	-465.140	0.862	0.820	0.090	-445.092	-450.230	-445.170	0.914	0.860	0.075
MoER-N	-464.139	-465.276	-455.123	0.841	0.790	0.100	-440.071	-445.130	-435.150	0.902	0.820	0.080
MR-STT	-582.211	-585.423	-575.323	0.735	0.710	0.150	-540.190	-545.312	-530.276	0.793	0.750	0.130
MR-SGLN	-570.175	-575.320	-565.244	0.724	0.700	0.160	-530.180	-535.240	-520.230	0.783	0.740	0.140
MR-ST	-561.148	-565.209	-555.230	0.743	0.720	0.155	-525.149	-530.210	-520.215	0.765	0.710	0.145
MR-SSN	-554.129	-555.201	-545.211	0.753	0.730	0.145	-520.135	-525.196	-515.180	0.777	0.690	0.150
MR-SN	-548.116	-545.198	-535.183	0.767	0.740	0.140	-510.118	-515.170	-505.195	0.781	0.680	0.155
MR-N	-533.103	-535.182	-525.170	0.771	0.750	0.135	-500.106	-505.130	-495.145	0.792	0.660	0.160

5.3 Predictive performance evaluation

To evaluate the predictive performance of the proposed MoER-SSMSN model, we conducted a Monte Carlo simulation study and compared it with widely used machine learning methods, including Random Forests, Gradient Boosting, and Neural Networks. In addition, the simulation design was constructed to investigate the robustness of the proposed heavy-tailed MoE framework against contamination and atypical observations, as well as its computational behavior in higher-dimensional settings. This experiment investigates how effectively the proposed model identifies mixture components when the dataset contains atypical or outlying observations. We generated 200 independent samples, each of size $n = 3000$, from the regression model in (3.1) with $g = 3$ mixture components. The error terms were drawn either from the normal mean variance Birnbaum-Saunders (NMVBS; Naderi et al., 2023) distribution or the extended skew normal (ESN; Capitanio et al., 2003) distribution. In particular, 200 replications were simulated from each family, using the two-component version of (3.1) with $\varepsilon_{ij} \sim \text{NMVBS}(0, \sigma_i^2, \lambda_i, \alpha_i)$, where $\alpha_1 = 1$ and $\alpha_2 = 2$, or $\varepsilon_{ij} \sim \text{ESN}(0, \sigma_i^2, \lambda_i, \nu_i)$ with $\nu_1 = 3$ and $\nu_2 = 5$. For both cases, equal mixing proportions $\pi_i = 1/2$ were assumed.

To further assess the scalability and computational stability of the proposed models

in moderately high-dimensional settings, the number of predictors was increased to $p = 10$. The covariate vector was defined as $\boldsymbol{x}_j = (1, x_{1j}, x_{2j}, \ldots, x_{10j})^\top$, where $x_{kj} \sim U(-1, 5)$, for $k = 1, \ldots, 10$. Two different configurations of regression coefficients were considered:

- **First scenario (S1):**

  $\boldsymbol{\beta}_1 = (-4, 4, 2, -1, 3, 0.5, -2, 1, 2.5, -1.5, 0.8)^\top$, $\boldsymbol{\beta}_2 = (0, 3, -2, 1, -1, 2, 1.5, -0.5, 3, -2, 1)^\top$,

  with $\sigma_1^2 = \sigma_2^2 = 0.4$.

- **Second scenario (S2):**

  $\boldsymbol{\beta}_1 = (-1, 3, 1.5, -2, 2, 0.5, -1, 2.5, -0.5, 1, 3)^\top$, $\boldsymbol{\beta}_2 = (3, -1, 2, 1, -2, 3, -1.5, 0.5, 2, -3, 1.5)^\top$,

  with $\sigma_1^2 = 0.6$ and $\sigma_2^2 = 0.9$.

The skewness parameters were fixed at $\lambda_1 = 2$ and $\lambda_2 = 3$. The presumed gating parameters were set as $\boldsymbol{\psi}_1 = (2, -3, 1, -1.5, 2.5, -0.5, 1, 2, -2, 0.5, -1)^\top$ and $\boldsymbol{\psi}_2 = (1, -2.5, 2, -1, 1.5, -2, 0.5, 1, -1.5, 2, -0.5)^\top$.

To evaluate robustness against contamination and outliers, an additional 150 observations were generated from $U(15, 25)$ and randomly assigned to one of the mixture components. This contamination mechanism was intentionally introduced to assess the stability and robustness properties of the proposed heavy-tailed MoER-SSMSN models under atypical data scenarios.

For each of the 200 replications, three benchmark MoER-SSMSN models (MoER-STT, MoER-SGLN, and MoER-ST) and three counterparts from the machine learning methods (Random Forests (RF), Gradient Boosting (GB), and Neural Networks (NN)) were fitted with $g = 2$ mixture components. All machine learning benchmark models were implemented in `R` using the corresponding packages and were evaluated on the same independent test sets as the proposed model. In each replication, the simulated data were randomly divided into training and testing sets with a ratio of 70% and 30%, respectively. The Random Forest model was fitted using the `randomForest` package with 500 trees and default splitting parameters. The Gradient Boosting model was implemented using the `gbm` package with 300 boosting iterations, a learning rate of 0.05, and interaction depth equal to 3. The feed-forward Neural Network was fitted using the `nnet` package with one hidden layer containing 10 neurons, using the logistic

Table 3: Performance comparison between the selected MoER-SSMSN models and machine learning methods under simulated data with outliers.

Error	Outlier	Measure	MoER-STT		MoER-SGLN		MoER-ST		RF		GB		NN	
			S1	S2	S1	S2	S1	S2	S1	S2	S1	S2	S1	S2
NMVBS	0	RMSE	1.22	1.26	1.21	1.27	1.23	1.26	1.18	1.24	1.18	1.24	1.18	1.25
		MAE	0.95	0.98	0.94	0.99	0.96	0.98	1.13	1.19	1.14	1.19	1.14	1.20
		R^2	0.956	0.710	0.961	0.710	0.957	0.711	0.962	0.716	0.959	0.713	0.957	0.710
		CPU Time (s)	13.42	14.85	14.58	14.94	13.97	14.11	2.34	2.48	2.75	2.81	3.12	3.25
		Avg. Iteration	238	241	240	242	235	237	–	–	–	–	–	–
	50	RMSE	1.41	1.46	1.38	1.47	1.40	1.46	1.59	1.63	1.57	1.64	1.58	1.64
		MAE	1.07	1.12	1.05	1.13	1.06	1.12	1.53	1.58	1.52	1.58	1.52	1.58
		R^2	0.819	0.597	0.807	0.595	0.815	0.597	0.832	0.605	0.820	0.601	0.818	0.599
		CPU Time (s)	17.14	17.48	17.31	17.56	16.82	16.95	2.61	2.74	2.94	3.08	3.41	3.55
		Avg. Iteration	344	347	345	348	341	343	–	–	–	–	–	–
ESN	0	RMSE	1.33	1.37	1.32	1.38	1.34	1.37	1.23	1.30	1.23	1.30	1.23	1.31
		MAE	1.00	1.04	0.99	1.05	1.01	1.04	1.17	1.24	1.18	1.25	1.18	1.25
		R^2	0.920	0.683	0.925	0.689	0.922	0.683	0.932	0.692	0.928	0.689	0.924	0.686
		CPU Time (s)	14.75	14.03	14.89	15.16	14.31	14.45	2.42	2.57	2.86	2.93	3.20	3.31
		Avg. Iteration	239	242	241	243	236	238	–	–	–	–	–	–
	50	RMSE	1.48	1.53	1.50	1.55	1.49	1.54	1.70	1.75	1.75	1.76	1.72	1.76
		MAE	1.13	1.18	1.15	1.20	1.14	1.19	1.64	1.69	1.69	1.71	1.66	1.70
		R^2	0.680	0.652	0.685	0.656	0.681	0.652	0.692	0.661	0.688	0.657	0.684	0.654
		CPU Time (s)	17.63	17.95	17.81	18.04	17.18	17.36	2.84	2.98	3.17	3.29	3.58	3.71
		Avg. Iteration	346	349	347	350	343	345	–	–	–	–	–	–

activation function and a maximum of 500 training iterations. These settings were fixed throughout the 200 simulation replications to ensure a consistent comparison among competing methods. The hyperparameter values were selected based on commonly used settings in the literature and fixed before conducting the simulation study. No additional regularization was applied in the neural network model.

For each simulated dataset with sample size $n = 3000$, the proposed MoER-SSMSN model, Random Forest, Gradient Boosting, and a feed-forward neural network were fitted. Predictive performance was assessed using RMSE, MAE, CPU Time (Seconds), and R^2 on an independent test set. In addition to predictive accuracy measures, computational aspects of the proposed estimation procedure were also examined through the average CPU time and the average number of iterations required for convergence. The simulation experiment was repeated 200 times, and the average values of these performance measures were reported.

The results reported in Table 3 provide a comprehensive comparison between the proposed MoER-SSMSN models and conventional machine learning approaches across

different error distributions, contamination levels, and parameter settings. Overall, the MoER-based models demonstrate competitive and often superior predictive performance, particularly in terms of RMSE and MAE, under both NMVBS and ESN error structures. In the absence of outliers, all models perform relatively well; however, the MoER-SGLN specification consistently achieves slightly lower RMSE and MAE values compared to MoER-STT and MoER-ST, indicating a better fit to asymmetric and heavy-tailed data. This advantage is especially noticeable in Scenario S1, where the separation between components is more pronounced. Although machine learning methods such as RF and GB exhibit strong R^2 values-occasionally exceeding those of the MoER models-their error metrics (RMSE and MAE) are generally higher, suggesting less precise predictions despite good overall variance explanation. NN tend to perform similarly to GB but do not outperform the MoER-based approaches in a consistent manner.

When contamination is introduced (150 outliers), the robustness of the MoER-SSMSN framework becomes more evident. The increase in RMSE and MAE remains comparatively moderate for the proposed heavy-tailed MoER models, indicating stronger resistance to outlying observations and contaminated data structures. In particular, MoER-SGLN maintains the best overall performance among the MoER variants, especially under the NMVBS distribution, where heavy tails and skewness are present. Under the ESN distribution, a similar trend is observed, although the gap between models slightly narrows. Machine learning methods experience a more substantial degradation in performance under contamination, with RF and NN showing the largest increases in error metrics. While R^2 values decrease for all models in the presence of outliers, the decline is less severe for the MoER-based models, suggesting more stable predictive behavior.

From a computational perspective, the reported CPU times and average iteration counts indicate that the proposed EM-type estimation procedure achieves stable convergence within a reasonable number of iterations, even in the presence of contamination and higher-dimensional covariates. These findings highlight the effectiveness of incorporating flexible skewed and heavy-tailed distributions within the mixture-of-experts framework, making the proposed MoER-SSMSN models particularly suitable for complex data scenarios involving heterogeneity and contamination.

6 Application to Economic Data

In this part of the study, we evaluate the performance of the proposed model by applying it to two real-world datasets and comparing the results with those obtained from the classic mixture (MR-N) and mixture of experts (MoER-N) models

6.1 Data sets

- Film rev. data set: The `film90` dataset, available in the `gamlss.data` R package, contains information on the revenues of films released in the year 1990. The data includes the box office revenue, measured in millions of USD, for each movie in the sample. The dataset also includes variables such as the genre, the production budget, and the number of screens the film was released on. This dataset is commonly used in the context of modeling financial data due to its diverse characteristics, including heavy-tailed distributions and significant skewness. The `film90` data provides a valuable real-world example for testing and comparing different statistical models, particularly those designed to handle non-normal distributions. We investigated the linear dependence of the `lboopen` variable on the `lnosc` and `lborevl` variables. Additionally, we considered the `lnosc` and `lborevl` variables as gating covariates.

- rent99 data set: The `rent99` dataset, available in the `gamlss.data` R package, provides information on rental prices in Munich for the year 1999. The dataset consists of the monthly rental prices, known as the net rent, which represents the amount left after deducting all operational expenses and additional costs, calculated on a per-square-meter basis. This data offers valuable insights into the housing market and is often used for modeling real estate prices. It is characterized by a skewed distribution and some degree of leptokurtosis, making it suitable for testing models that handle non-normal data. The `rent99` data is commonly employed for illustrating the application of statistical techniques in economic and financial modeling. We investigated the linear dependence of the `rent` variable on the `rentsqm` variable. Additionally, we considered the `rentsqm` variable as gating covariate.

Table 4: Selected component, BIC values, p-values, and rankings for the models applied to the rent99 and film rev. datasets.

Models	rent99					film rev.				
	BIC	g	KS	p-value	Rank	BIC	g	KS	p-value	Rank
MoER-N	14341.46	3	0.069	0.15	15	18906.71	3	0.075	0.22	15
MoER-SN	14205.71	2	0.058	0.55	10	18832.61	2	0.064	0.63	11
MoER-ST	14152.61	2	0.009	0.96	4	18784.49	2	0.011	0.97	6
MoER-STN	14147.49	2	0.009	0.98	3	18775.35	2	0.011	0.98	3
MoER-STT	14131.35	2	0.006	0.99	1	18760.76	2	0.010	0.99	2
MoER-SGLN	14135.76	2	0.006	0.99	2	18751.92	2	0.010	0.98	1
MoER-SSN	14166.11	2	0.015	0.91	7	18783.78	2	0.018	0.95	5
MoER-MMNEH	14158.42	2	0.011	0.95	5	18781.64	2	0.013	0.96	4
MoER-MMNE	14163.18	2	0.013	0.93	6	18788.51	2	0.015	0.95	7
MR-N	14450.82	3	0.077	0.12	16	18955.56	3	0.089	0.15	16
MR-SN	14319.56	2	0.063	0.42	14	18869.08	2	0.075	0.45	14
MR-ST	14288.08	2	0.018	0.81	13	18840.82	2	0.019	0.79	12
MR-STN	14232.82	2	0.015	0.89	11	18820.56	2	0.022	0.84	10
MR-STT	14197.56	2	0.012	0.90	8	18809.08	2	0.018	0.90	9
MR-SGLN	14203.08	2	0.019	0.93	9	18797.56	2	0.022	0.89	8
MR-SSN	14252.23	2	0.026	0.83	12	18845.34	2	0.041	0.86	13

6.2 Experiments and Evaluation

In every trial, models from the MoER-SSMSN and MR-SSMSN families, together with the MoER-MMNE and MoER-MMNEH models proposed by Sepahdar et al. (2022), were fitted to the data for varying values of g, with the optimal number of mixture components selected according to the BIC criterion. The first experiment focused on evaluating the effectiveness of the proposed models for clustering tasks, particularly in the presence of regression relationships among the variables. A range of models was evaluated on the `rent99` and `film rev.` datasets, with g values spanning from 1 to 5. Table 4 shows the significant results, namely BIC scores, component counts (g), KS statistics, and p-values for all models as applied to the datasets.

As shown in Table 4, the skewed models consistently outperform their symmetric counterparts across both datasets. Specifically, when analyzing the `rent99` dataset, a substantial improvement in both BIC and KS statistics is observed with the MoER-STT model. Every skewed model opted for a two-component structure, and its performance

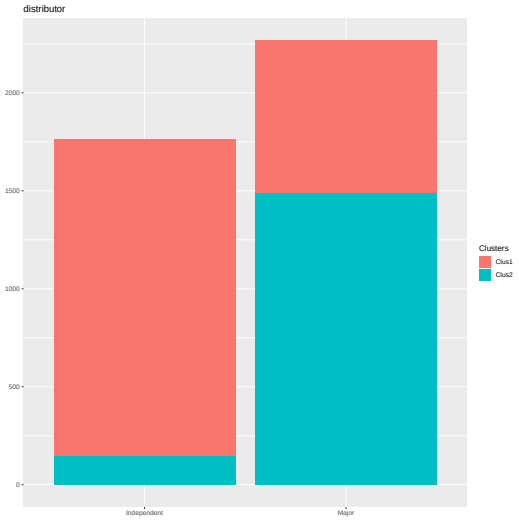

Figure 2: Bar chart of categorical attribute, in the `film rev.` data set.

was better than that of the MoER-N and MR-N models, which used three components. Additionally, the asymmetric MoER-SSMSN models generally demonstrated superior fitting performance compared to the MoER-N model. Notably, the MoER-SGLN, a two-component heavy-tailed model, yielded the best results with a BIC of 18751.92. Figures 2 and 3 are obtained based on the classification results derived from the MoER-STT and MoER-SGLN models, respectively.

Figure 2 suggests that most films distributed by independent distributors tend to have lower values for both the log of box office opening revenues and the log of box office revenues after the first week. In contrast, films released by major distributors generally exhibit higher revenue levels on these two measures. However, it should also be noted that some films associated with major distributors still show relatively low opening-week and post–first-week box office revenues.

Using Figure 3, based on the monthly net rent per month, Almost all houses located in top locations belong to the second class, and the monthly rental prices of these houses are generally higher than this of the other class. Also, The presence of a standard bathroom and kitchen has a significant positive effect on increasing monthly rental prices. Almost all cars in the cluster 2 have high the monthly rental prices. Houses equipped with central heating generally have higher rental prices and mostly belong to the cluster 2. This may be explained by the fact that central heating reduces energy costs, which play an important role in household economics in Europe. An

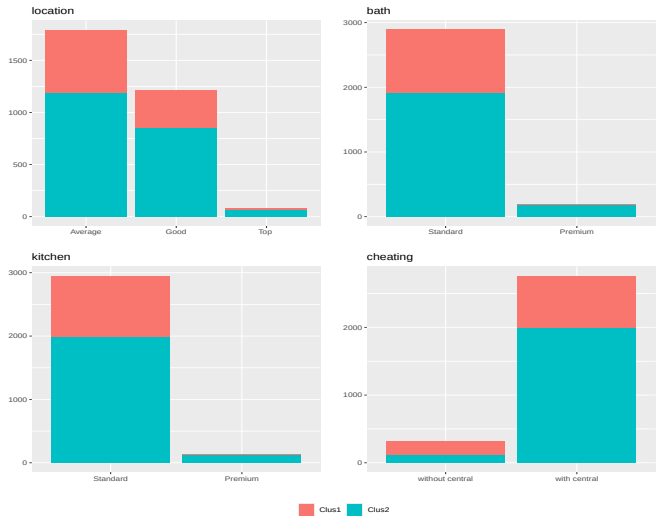

Figure 3: Bar charts of all categorical attributes in the `rent99` data set.

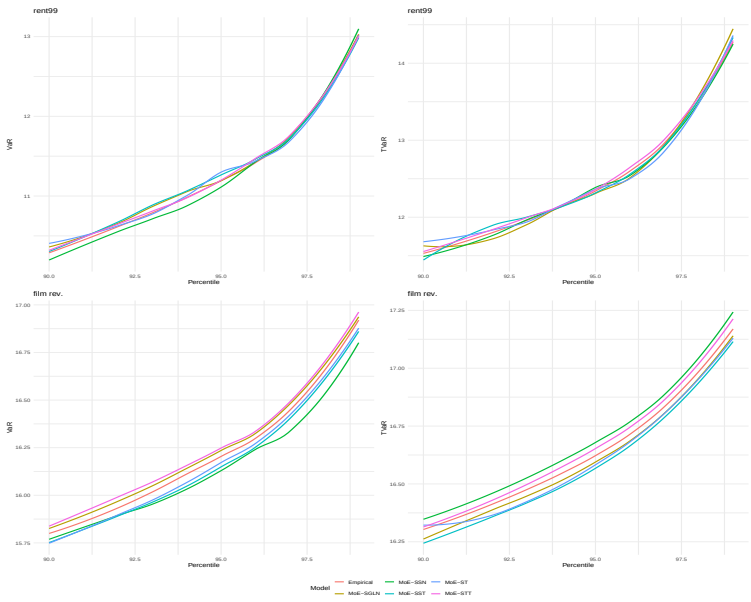

Figure 4: Plots of VaR (left) and TVaR (right) against varying confidence levels (q) for the rent99 (top) and Film rev. (bottom) datasets.

evaluation of the precision of the forecasted VaR and TVaR values will be conducted, drawing upon the parameter estimates from the prior section. For this evaluation, three million samples are generated from each model. Please note that VaR aligns with

the q-th percentile of the loss simulation data, whereas TVaR signifies the average loss, consistently exceeding VaR. Table 5 displays the detailed VaR and TVaR estimations across different models and four datasets, with confidence levels of 99%, 97.5%, and 95%. The results clearly indicate that the MoER-SSMSN distributions produce VaR and TVaR predictions that are much closer to the true values in most cases. Figure 4 provides a visual comparison of empirical values and theoretical VaR/TVaR predictions from the five best models, with confidence levels from 90% to 99%. Projected risk measures closely resemble those observed in reality.

Table 5: Evaluation of VaR and TVaR estimates for two real-world economic datasets, based on the MoER-SSMSN sub-models, across confidence levels of 99%, 97.5%, and 95%.

Data	Criteria	Level	Empir.	MoE-N	MoE-SN	MoE-ST	MoE-STN	MoE-STT	MoE-SGLN	MoE-SSN
rent99	VaR	0.99	13.077	15.520	13.493	13.065	13.069	13.074	13.072	13.050
		0.975	11.947	14.005	12.383	12.110	12.052	11.988	12.008	12.305
		0.95	11.222	11.830	10.995	11.158	11.179	11.201	11.190	11.013
	TVaR	0.99	14.305	18.096	16.639	14.243	14.250	14.290	14.273	14.197
		0.975	13.167	15.361	13.983	13.205	13.191	13.165	13.175	13.251
		0.95	12.339	13.480	13.015	12.402	12.377	12.355	12.364	12.456
Film rev.	VaR	0.99	16.920	18.150	17.122	17.104	16.859	16.887	16.895	16.806
		0.975	16.523	17.052	16.362	16.401	16.512	16.520	16.524	16.486
		0.95	16.210	15.710	15.930	16.015	16.193	16.202	16.214	16.170
	TVaR	0.99	17.172	19.411	17.530	17.272	17.249	17.226	17.205	17.261
		0.975	16.902	18.236	17.140	16.996	16.920	16.917	16.905	16.932
		0.95	16.622	17.430	16.901	16.840	16.649	16.640	16.625	16.670

7 Concluding Thoughts

In this study, we present a novel robust model based on the mixture of linear experts approach, tailored for handling heavy-tailed data. The proposed MoER-SSMSN model leverages the scale-shape mixture of skew-normal (SSMSN) distributions to effectively address key challenges such as accommodating heavy-tailed distributions and managing outliers. By extending the classic MoER framework, this model provides a versatile and powerful tool that adapts to the complexities of real-world data. One of the key strengths of the MoER-SSMSN model is its flexibility, which allows it to generalize the traditional MoER approach. A significant advantage of this model is its ability

to incorporate covariates into the gating function, leading to enhanced data classification and improved model performance. We have also introduced an innovative Expectation-Maximization (EM)-type algorithm for efficiently estimating the model parameters using the hierarchical structure of the SSMSN distribution. This algorithm is implemented within the R programming environment, offering a practical and effective solution for researchers and practitioners. To assess the performance of the MoER-SSMSN model, we conducted four Monte Carlo simulation studies, focusing on tasks such as nonlinear regression, prediction, and model-based clustering. The results of these simulations demonstrate the model's robustness against outliers and unusual observations, confirming its suitability for handling complex data structures. Additionally, a real-world case study illustrated the model's applicability and its advantages in practical scenarios.

Looking ahead, one promising avenue for future research involves the development of variable selection techniques for both the regression and gating components of the model. We also plan to explore the extension of this model from univariate to multivariate regression settings, which is currently under investigation. For further reading, we direct attention to the work of Sepahdar et al. (2022), who proposed an exact ECM algorithm for mean mixture normal distributions. Another exciting direction is the adoption of a fully Bayesian framework for model inference and prediction, a concept explored in earlier works (Peng et al., 1996; Zens and Lerch, 2019).

Acknowledgments

The author would like to thank the Editor and the anonymous reviewers for their valuable comments and constructive suggestions, which helped improve the quality and presentation of this manuscript.

Declarations

Conflict of interest: The authors declare that they have no Conflict of interest.
Competing Interests: The author declares that there are no competing interests.
Funding: The author declares that no funding was received for this research.

References

Azzalini A. A Class of Distributions Which Includes the Normal Ones. Scandinavian Journal of Statistics. 1985;12(2):171–178.

Azzalini A. Further Results on a Class of Distributions Which Includes the Normal Ones. Statistica. 1986;46:199–208.

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. 2003;65(2):367–389.

Azzalini A, Dalla Valle A. The Multivariate Skew-Normal Distribution. Biometrika. 1996;83(4):715–726.

Bai Z, Rao CZ, Wu CZ. Model Selection: The Efficient Determination Criterion. Journal of Multivariate Analysis. 1989;30(2):180–198.

Basford K, Greenway D, McLachlan G, Peel D. Standard errors of fitted component means of normal mixtures. Computational Statistics. 1997;12(1):1–18.

Baudry JP, Celeux G. EM for mixtures: convergence, likelihood monotonicity, and initialization. Statistics and Computing. 2015;25(4):713–728.

Biernacki C, Celeux G, Govaert G. Assessing a Mixture Model for Clustering with the Integrated Completed Likelihood. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2000;22(7):719–725.

Capitanio A, Azzalini A, Stanghellini E. Graphical models for skew-normal variates. Scandinavian Journal of Statistics. 2003;30(1):129–144.

Chamroukhi F. Skew t mixture of experts. Neurocomputing. 2017;266:390–408.

Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society: Series B. 1977;39(1):1–38.

Gormley IC, Murphy TB. A mixture of experts model for rank data with applications in election studies. 2008;.

Hubert L, Arabie P. Comparing Partitions. Journal of Classification. 1985;2(1):193–218.

Jacobs RA, Jordan MI, Nowlan SJ, Hinton GE. Adaptive mixtures of local experts. Neural Computation. 1991;3(1):79–87.

Jamalizadeh A, Lin TI. A general class of scale-shape mixtures of skew-normal distributions: properties and estimation. Computational Statistics. 2017;32(2):451–474.

Liu C, Rubin DB. The ECME Algorithm: A Simple Extension of EM and ECM with Faster Monotone Convergence. Biometrika. 1994;81(4):633–648.

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.

Mazza A, Punzo A. Model-based clustering and classification with mixtures of contaminated Gaussian regression models. Advances in Data Analysis and Classification. 2017;11:753–777.

Mazza A, Punzo A. Mixtures of multivariate contaminated normal regression models. Statistical Papers. 2020;61(2):787–822.

McNeil AJ, Frey R, Embrechts P. Quantitative risk management: concepts, techniques and tools-revised edition. Princeton university press; 2015.

Meilă M. Comparing Clusterings—An Information Based Distance. Journal of Multivariate Analysis. 2007;98(5):873–895.

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

Naderi M, Mirfarah E, Wang WL, Lin TI. Robust mixture regression modeling based on the normal mean-variance mixture distributions. Computational Statistics & Data Analysis. 2023;180:107661.

Nguyen HD, McLachlan GJ. Laplace mixture of linear experts. Computational Statistics & Data Analysis. 2016;93:177–191.

Peng F, Jacquier E, Strahan P. Bayesian Inference for Mixture Models. Journal of the American Statistical Association. 1996;91(433):187–196.

Poursadeghfard T, Nematollahi A, Jamalizadeh A. The Extended Birnbaum–Saunders Distribution Based on the Scale Shape Mixture of Skew Normal Distributions. Journal of Statistical Theory and Applications. 2021;20:481–517.

Rogers WH, Tukey JW. Understanding some long-tailed symmetrical distributions. Statistica Neerlandica. 1972;26(3):211–226.

Santana L, Vilca F, Leiva V. Influence analysis in skew-Birnbaum–Saunders regression models and applications. Journal of Applied Statistics. 2011;38(8):1633–1649.

Schwarz G. Estimating the Dimension of a Model. Annals of Statistics. 1978;6(2):461–464.

Sepahdar A, Madadi M, Balakrishnan N, Jamalizadeh A. Parsimonious mixture-of-experts based on mean mixture of multivariate normal distributions. Stat. 2022;11(1):e421.

Tanimoto T. An Elementary Mathematical Theory of Classification and Prediction. IBM Internal Report. 1958;Originally published as an internal technical report, IBM Corporation.

Wang P, Puterman ML, Cockburn I, Le N. Mixed Poisson regression models with covariate dependent rates. Biometrics. 1998;54(1):237–247.

Zeller CB, Cabral CR, Lachos VH. Robust mixture regression modeling based on scale mixtures of skew-normal distributions. Test. 2016;25(2):375–396.

Zens M, Lerch S. Probabilistic Forecasting and Bayesian Model Averaging for Multivariate Data. Bayesian Analysis. 2019;14(3):945–970.

Appendix

A Particular instances of SSMSN distributions

A.1 The STT distribution

Consider two independent random variables, τ_1 and τ_2, each of which follows a gamma distribution with identical shape and rate parameters of $\frac{\nu_i}{2}$, where $i = 1, 2$. Specifically,

we have: $\tau_i \sim \Gamma(\nu_i/2, /\nu_i/2)$ for $i = 1, 2$. The joint pdf of τ_1 and τ_2 is given by the product of their individual gamma distributions:

$$h(\boldsymbol{\tau}; \boldsymbol{\nu}) = f_\Gamma\left(\tau_1; \frac{\nu_1}{2}, \frac{\nu_1}{2}\right) f_\Gamma\left(\tau_2; \frac{\nu_2}{2}, \frac{\nu_2}{2}\right), \tag{A.1}$$

where $\boldsymbol{\tau} = (\tau_1, \tau_2)^\top$ and $\boldsymbol{\nu} = (\nu_1, \nu_2)^\top$, and f_Γ indicates the pdf of the gamma distribution. Thus, the random variable Y, as defined in (2.2), is described as following a skew-TT distribution with parameters $(\mu, \sigma^2, \lambda, \nu_1, \nu_2)$, indicated as $Y \sim STT(\mu, \sigma^2, \lambda, \nu_1, \nu_2)$. The pdf for the skew-$TT$ distribution is expressed as:

$$f_{\text{STT}}(y; \mu, \sigma^2, \lambda, \nu_1, \nu_2) = \frac{2}{\sigma} t(u; \nu_1)\, T(\lambda u; \nu_2), \quad y, \sigma, \nu_1, \nu_2 > 0, \quad y, \mu, \lambda \in \mathbb{R}, \tag{A.2}$$

where $t(\cdot; \nu_1)$ and $T(\cdot; \nu_2)$ represent the pdf and cdf of the t distribution with ν_1 and ν_2 degrees of freedom, respectively.

Remark A.1. If $\nu_1 = \nu_2 = \nu$, then $Y \sim \text{STT}(\mu, \sigma^2, \lambda, \nu_1, \nu_2)$, which we denote by $Y \sim \text{STT}(\alpha, \eta, \lambda, \nu)$.

It should be noted that the STT class contains several well-known distributions as special cases. Some important members are presented below.

(I) The distribution is derived by setting $\tau_1 \sim \Gamma\left(\frac{\nu}{2}, \frac{\nu}{2}\right)$ and $\tau_2 = 1$ with probability 1. Given these conditions, the distribution that results is known called the skew-t normal (STN). The corresponding pdf is expressed as:

$$f_{\text{STN}}(y; \mu, \sigma^2, \lambda, \nu) = \frac{2}{\sigma} t(u; \nu)\, \Phi(\lambda u), \quad \sigma, \nu > 0,\ y, \mu\lambda \in \mathbb{R}. \tag{A.3}$$

(II) The distribution emerges by setting $\tau_1 = 1$ with probability 1 and $\tau_2 \sim \Gamma\left(\frac{\nu}{2}, \frac{\nu}{2}\right)$. Under these conditions, the distribution that results is known as the skew normal-t (SNT). Its pdf is given by:

$$f_{\text{SNT}}(y; \mu, \sigma^2, \lambda, \nu) = \frac{2}{\sigma} \phi(u)\, T(\lambda u; \nu)\, \xi_1(y; \alpha, \eta), \quad \sigma, \nu > 0,\ y, \mu, \lambda \in \mathbb{R}. \tag{A.4}$$

(III) The skew normal (SN) distribution arises when both τ_1 and τ_2 are set to 1 with probability 1. Under these conditions, the pdf is given by:

$$f_{\text{SN}}(y; \mu, \sigma^2, \lambda) = \frac{2}{\sigma} \phi(u)\, \Phi(\lambda u)\, \xi_1(y; \alpha, \eta), \quad \sigma, \nu > 0,\ y, \mu, \lambda \in \mathbb{R}. \tag{A.5}$$

For further details, see Poursadeghfard et al. (2021). From (2.7), the conditional pdfs of τ_1 and τ_2 given $Y = y$ can be expressed as

$$\tau_1 \mid Y = y \sim \Gamma\left(\frac{\nu_1 + 1}{2}, \frac{\nu_1 + u^2}{2}\right), \tag{A.6}$$

and

$$f_{\tau_2|Y}(\tau_2 \mid Y = y) = \frac{\frac{\nu_2}{2}^{\nu_2/2}}{\Gamma\left(\frac{\nu_2+1}{2}\right)} \tau_2^{\frac{\nu_2}{2}-1} \exp\left(-\frac{\nu_2\tau_2}{2}\right) \frac{\Phi\left(\sqrt{\tau_2}\, u\right)}{T\left(u;\nu_2\right)}, \tag{A.7}$$

where distribution of $\tau_2 \mid Y = y$ is weighted gamma distribution.

Accordingly, the corresponding conditional expectations are given by

$$\mathrm{E}[\tau_1 \mid Y = y] = \frac{\nu_1 + 1}{\nu_1 + u^2},\ \mathrm{E}[\log \tau_1 \mid Y = y] = \mathrm{DG}\left(\frac{\nu_1 + 1}{2}\right) - \log\left(\frac{\nu_1 + u^2}{2}\right),$$

$$\mathrm{E}[\tau_2 \mid Y = y] = \frac{T\left(\lambda u \sqrt{\frac{\nu_2+2}{\nu_2}}; \nu_2 + 2\right)}{T\left(\lambda u; \nu_2\right)},$$

$$\begin{aligned}\mathrm{E}[\log \tau_2 \mid Y = y] = &\frac{T\left(\lambda u \sqrt{\frac{\nu_2+2}{\nu_2}}; \nu_2 + 2\right)}{T\left(\lambda u; \nu_2\right)} - \log\left(\frac{\nu_2}{2}\right) - 1 + \mathrm{DG}\left(\frac{\nu_2 + 1}{2}\right) \\ &+ \frac{1}{T\left(\lambda u; \nu_2\right)} \int_{-\infty}^{\lambda u} \frac{x^2 - 1}{\nu_2 + x^2}\, t(x; \nu_2)\, dx,\end{aligned}$$

$$\begin{aligned}\mathrm{E}[\gamma\tau_2 \mid Y = y] = &\frac{1}{T\left(\lambda u; \nu_2\right)} \Bigg[\lambda u\, T\left(\lambda u \sqrt{\frac{\nu_2 + 2}{\nu_2}}; \nu_2 + 2\right) \\ &+ \frac{\Gamma\left(\frac{\nu_2+1}{2}\right)}{\Gamma\left(\frac{\nu_2}{2}\right)} \left(\frac{\nu_2}{2}\right)^{1/2} \left(\nu_2 + \lambda^2 u^2\right)^{-\frac{\nu_2+1}{2}} \Bigg],\end{aligned} \tag{A.8}$$

where $\mathrm{DG}(x) = \Gamma'(x)/\Gamma(x)$ denotes the digamma function.

A.2 The ST distribution

Consider two independent random variables, $\tau_1 = \tau_2 = \tau$, which follows a gamma distribution with identical shape and rate parameters of $\frac{\nu}{2}$. The pdf of $\tau \sim \Gamma\left(\frac{\nu}{2}, \frac{\nu}{2}\right)$ is given by

$$h(\tau; \nu) = f_\Gamma\left(\tau; \frac{\nu}{2}, \frac{\nu}{2}\right). \tag{A.9}$$

Consider the random variable Y as defined in (2.2). Under certain conditions, it follows the skew-t distribution with parameters $(\mu, \sigma^2, \lambda, \nu)$. This is denoted as $Y \sim$

$\text{ST}(\mu, \sigma^2, \lambda, \nu)$. The corresponding pdf is given by:

$$f_{\text{ST}}(y; \mu, \sigma^2, \lambda, \nu) = \frac{2}{\sigma} t\,(u; \nu)\, T\left(\lambda u \sqrt{\frac{\nu+1}{u^2+\nu}}; \nu+1\right) \xi_1(y; \alpha, \eta), \quad \sigma, \nu > 0,\ y, \mu, \lambda \in \mathbb{R}. \tag{A.10}$$

From Equation (2.7), the conditional pdf of τ given $Y = y$ can be expressed as

$$f(\tau \mid Y = y) = \frac{\tau^{\frac{\nu+1}{2}-1} \exp\left\{-\frac{\tau}{2}\left[u^2+\nu\right]\right\}}{\Gamma\left(\frac{\nu+1}{2}\right)\left[\frac{u^2+\nu}{2}\right]^{-\frac{\nu+1}{2}}} \frac{\Phi\left(\sqrt{\tau}\, u\right)}{T\left(\lambda u \sqrt{\frac{\nu+1}{u^2+\nu}}; \nu+1\right)}. \tag{A.11}$$

Accordingly, the conditional expectations of τ and $\log \tau$ given $Y = y$ are given by

$$E(\tau \mid Y = y) = \frac{\nu+1}{u^2+\nu} \frac{T\left(\lambda u \sqrt{\frac{\nu+3}{u^2+\nu}}; \nu+3\right)}{T\left(\lambda u \sqrt{\frac{\nu+1}{u^2+\nu}}; \nu+1\right)}, \tag{A.12}$$

and

$$\begin{aligned} E(\log \tau \mid Y = y) = {} & \text{DG}\left(\frac{\nu+1}{2}\right) - \log\left(\frac{u^2+\nu}{2}\right) \\ & + \frac{\nu+1}{u^2+\nu}\left[\frac{T\left(\lambda u \sqrt{\frac{\nu+3}{u^2+\nu}}; \nu+3\right)}{T\left(\lambda u \sqrt{\frac{\nu+1}{u^2+\nu}}; \nu+1\right)} - 1\right] \\ & + \frac{\lambda u\,[u^2-1]}{\nu(\nu+1)[u^2+\nu]^3} \frac{t\left(\lambda u \sqrt{\frac{\nu+1}{u^2+\nu}}; \nu+1\right)}{T\left(\lambda u \sqrt{\frac{\nu+1}{u^2+\nu}}; \nu+1\right)} \\ & + \frac{1}{T\left(\lambda u \sqrt{\frac{\nu+1}{u^2+\nu}}; \nu+1\right)} \int_{-\infty}^{\lambda u \sqrt{\frac{\nu+1}{u^2+\nu}}} g(x; \nu)\, t(x; \nu+1)\, dx, \end{aligned} \tag{A.13}$$

where the function $g(x; \nu)$ is defined as

$$g(x; \nu) = \text{DG}\left(\frac{\nu+2}{2}\right) - \text{DG}\left(\frac{\nu+1}{2}\right) - \log\left(\frac{1+x^2}{\nu+1}\right) + \frac{(\nu+1)x^2 - (\nu-1)}{(\nu+1)(\nu+1+x^2)}.$$

Furthermore, the conditional expectation of ξ given $Y = y$ can be expressed as

$$E(\gamma\tau \mid Y = y) = \lambda u E(\tau \mid Y = y) + \frac{\Gamma\left(\frac{\nu+2}{2}\right)}{\sqrt{\pi}\Gamma\left(\frac{\nu+1}{2}\right)} \frac{\left(\frac{u^2+\nu}{\lambda^2 u^2 + u^2 + \nu}\right)^{\frac{\nu+1}{2}}}{T\left(\lambda u \sqrt{\frac{\nu+1}{u^2+\nu}}, \nu+1\right)}. \tag{A.14}$$

A.3 The SGLN Distribution

If Y complies with a generalized Laplace (GL) distribution, it is indicated as $Y \sim \text{GL}(\mu, \sigma^2, \nu)$. The pdf of Y is presented as:

$$f_{\text{GL}}(y; \mu, \sigma^2, \nu) = \frac{1}{\sigma\sqrt{2\pi}2^{\nu-1}\Gamma(\nu)} |u|^{\nu-\frac{1}{2}} K_{\nu-\frac{1}{2}}(|u|),$$

where $K_r(x)$ represents the modified Bessel function. Now, assume $\tau_1 \sim \frac{1}{\chi^2_{2\nu}}$ and $\tau_2 = 1$ with probability 1, as described in (2.3), where $\chi^2_{2\nu}$ denotes the chi-squared distribution with 2ν degrees of freedom. Under these conditions, the random variable Y is said to follow a skew generalized Laplace normal (SGLN) distribution. The pdf of this distribution is given by:

$$f_{\text{SGLN}}(y; \mu, \sigma^2, \lambda, \nu) = 2\, f_{\text{GL}}(y; \mu, \sigma^2, \nu)\, \Phi(\lambda u),$$

and we denote this distribution as $Y \sim \text{SGLN}(\mu, \sigma^2, \lambda, \nu)$.

Remark A.2. Several notable instances of this distribution are as follows:

(I) If $\tau_1 = 1$ with probability one and $\tau_2 \sim \frac{1}{\chi^2_{2\nu}}$, then

$$Y \sim \text{SNGL}(\mu, \sigma^2, \lambda, \nu), \quad f_{\text{SNGL}}(y) = \frac{2}{\sigma}\, \phi(u)\, F_{\text{GL}}(\lambda u; 0, 1, \nu).$$

(II) If $\tau_1 \sim \frac{1}{\chi^2_{2\nu_1}}$ and $\tau_2 \sim \Gamma\left(\frac{\nu_2}{2}, \frac{\nu_2}{2}\right)$, then

$$Y \sim \text{SGLT}(\mu, \sigma^2, \lambda, \nu_1, \nu_2), \quad f_{\text{SGLT}}(y) = 2\, f_{\text{GL}}(y; \mu, \sigma^2, \nu_1)\, T(\lambda u; 0, 1, \nu_2).$$

(III) If $\tau_1 \sim \Gamma\left(\frac{\nu_1}{2}, \frac{\nu_1}{2}\right)$ and $\tau_2 \sim \frac{1}{\chi^2_{2\nu_2}}(2\nu_2)$, then

$$Y \sim \text{STGL}(\mu, \sigma^2, \lambda, \nu_1, \nu_2), \quad f_{\text{STGL}}(y) = 2\, t(y; \mu, \sigma^2, \nu_1)\, F_{\text{GL}}(\lambda u; 0, 1, \nu_2),$$

where $F_{\text{GL}}(\cdot; \xi, \omega^2, \nu_2)$ denotes the cdf of the GL distribution.

From (2.7), the conditional distribution of ξ given $Y = y$ is

$$\tau \mid Y = y \sim \text{GIG}\left(\frac{1}{2} - \nu, 1, u^2\right),$$

where GIG denotes the Generalized Inverse Gaussian distribution (see Jørgensen [21]). Therefore,

$$E(\tau \mid Y = y) = |u|^{-1} \frac{K_{\nu+1}(|u|)}{K_\nu(|u|)}, \quad E(\log \tau \mid Y = y) = -\log|u| + \frac{K'_\nu(|u|)}{K_\nu(|u|)}. \tag{A.15}$$

A.4 The SSL distribution

Consider the scenario where $\tau_1 \sim \text{Beta}\left(\frac{\nu}{2}, 1\right)$ and $\tau_2 = 1$ with probability one, as outlined in equation (2.2). Under these assumptions, the random variable Y follows the skew slash normal (SSN) distribution, denoted as $Y \sim \text{SSN}(\lambda, \nu)$.

The pdf of Y is given by:

$$f_{\text{SSN}}(y; \lambda, \nu) = 2\, f_s(y; \nu)\, \Phi(\lambda y),$$

where $f_s(\cdot; \nu)$ represents the pdf of the slash distribution with shape parameter ν, which is defined as:

$$f_s(y; \mu, \sigma^2, \nu) = \begin{cases} \frac{\nu\, \Gamma\left(\frac{\nu+1}{2}\right) 2^{\frac{\nu}{2}-1} G\left(\frac{u^2}{2}; \frac{\nu+1}{2}\right)}{\sigma \sqrt{\pi} |u|^{\nu+1}}, & \text{if } y \neq \mu, \\ \frac{\nu}{\sigma \sqrt{2\pi}(\nu+1)}, & \text{if } y = \mu, \end{cases}$$

where $G(\cdot; \nu)$ represents the cdf of the gamma distribution with scale parameter 1 and shape parameter ν, as explained in Rogers and Tukey (1972). From (2.7), the conditional pdf of τ given $Y = y$ is

$$f(\tau \mid Y = y) = \begin{cases} \frac{|u|^{\nu+1} \tau^{\frac{\nu-1}{2}} \exp\left(-\frac{1}{2}u^2\right)}{\Gamma\left(\frac{\nu+1}{2}\right) 2^{\frac{\nu+1}{2}} G\left(\frac{u^2}{2}; \frac{\nu+1}{2}\right)}, & y \neq \mu, \\ \frac{\nu+1}{2} \tau^{\frac{\nu+1}{2}-1}, & y = \mu. \end{cases}$$

The conditional expectations of τ and $\log \tau$ given $Y = y$ are respectively

$$E(\tau \mid Y = y) = \begin{cases} \frac{(\nu+1) G\left(u^2; \frac{\nu+3}{2}\right)}{u^2 G\left(\frac{u^2}{2}; \frac{\nu+1}{2}\right)}, & y \neq \mu, \\ \frac{\nu+1}{\nu+3}, & y = \mu, \end{cases}$$

and

$$E(\log \tau \mid Y = y) = \begin{cases} \log\left(\frac{2}{u^2}\right) + \frac{\int_0^{u^2/2} (\log x)\, x^{\frac{\nu-1}{2}} e^{-x}\, dx}{\Gamma\left(\frac{\nu+1}{2}\right) G\left(\frac{u^2}{2}; \frac{\nu+1}{2}\right)}, & y \neq \mu, \\ -\frac{2}{\nu+1}, & y = \mu. \end{cases} \tag{A.16}$$