Robust Bayesian Inference for the Censored Mixture of Experts Model Using Heavy-Tailed Distributions

Elham Mirfarah,Mehrdad Naderi, Tsung-I Lin, Wan-Lun Wang

MTH422 Course Project

Ahana Bose (231080009)
Sneha Karmakar (231080087)
Sohini Bhadra (231080089)
Rishikesh Dargad (220891)

Course Instructor : Prof. Arnab Hazra

20 April 2025

Contents

1 Introduction **3**

2 Model specification and Bayesian inference **3**
2.1 Notation and background material . . . 3
2.2 Model formulation . . . 3
2.3 Priors and Hyper-parameter Specifications . . . 4
2.4 Full Conditional Posteriors . . . 5
2.5 Sampling Strategy for the Indexing Parameter . . . 6
2.5.1 Slash (SL) Distribution . . . 6
2.5.2 Contaminated Normal (CN) Distribution . . . 7
2.5.3 Student's t Distribution . . . 7
2.5.4 Variance-Gamma (VG) Distribution . . . 8
2.5.5 Tail-Inflated Normal (TIN) Distribution . . . 8
2.5.6 Laplace and Normal Distributions . . . 8

3 Simulations **9**
3.1 Comparison of MH Strategies for Sampling Indexing Parameters in MoE-T-CR and MoE-VG-CR Models . . . 9
3.1.1 Objective . . . 9
3.1.2 Methodology . . . 9
3.1.3 MCMC Sampling . . . 9
3.1.4 Accuracy Metrics . . . 9
3.2 Interpretation . . . 10
3.2.1 Equal Variance Scenario . . . 10
3.2.2 Unequal Variance Scenario . . . 11
3.2.3 Hierarchical vs. Non-Hierarchical Modeling . . . 11
3.2.4 Prior Sensitivity . . . 11
3.3 Comparison of Bayesian and EM-based Inferences . . . 11
3.4 Interpretation . . . 12
3.5 Sensitivity Analysis of Parameter Estimation in the Presence of Outliers . . . 12
3.5.1 Interpretation . . . 15

4 Case study: Wage Data **15**
4.1 Model Interpretation and Comparison 15
4.2 Plots 18
4.2.1 ACF plots 18
4.2.2 Posterior Density Plots 19
4.2.3 Trace Plots 20

5 Conclusion **21**

6 Acknowledgment **21**

7 Our Contribution **21**

1 Introduction

Censored data, where responses are only partially observed, frequently occurs in fields like biomedicine, clinical trials, and econometrics. Traditional regression models often assume normally distributed errors, making them sensitive to outliers and latent heterogeneity, which can lead to poor estimation and classification performance.

The Mixture of Experts (MoE) model addresses heterogeneity by dividing data into subgroups, each with its own regression structure. However, standard MoE models still rely on normality assumptions and are not well-equipped to handle censoring or heavy-tailed distributions.

To overcome these limitations, this study proposes a **Bayesian Mixture of Experts (MoE)** model with **Scale Mixture of Normal (SMN)** errors, allowing for more robust handling of outliers and extreme values. The model, called **MoE-SMN-CR**, supports left, right, and interval censoring. It uses **Ultimate Pólya-Gamma (UPG)** augmentation to efficiently estimate gating parameters and perform posterior inference.
Simulation studies and real data analysis demonstrate that the proposed Bayesian approach improves estimation accuracy, classification performance, and robustness compared to traditional likelihood-based methods.

2 Model specification and Bayesian inference

2.1 Notation and background material

Throughout this paper, we use the following notations:

- $\phi(\cdot;\mu,\sigma^2)$ and $\Phi(\cdot;\mu,\sigma^2)$ denote the probability density function (pdf) and cumulative distribution function (cdf) of the normal distribution $\mathcal{N}(\mu,\sigma^2)$, respectively.
- Gamma(α,η) is the gamma distribution with mean α/η.
- IG(α,η) represents the inverse-gamma distribution with mean $\eta/(\alpha-1)$.
- TN$(\mu,\sigma^2;(a,b))$ and TG$(\alpha,\eta;(a,b))$ denote the truncated normal and truncated gamma distributions on the interval (a,b), respectively.
- $\mathcal{U}(a,b)$ is the uniform distribution over the interval (a,b).
- BE(α,η) is the beta distribution with mean $\alpha/(\alpha+\eta)$.

The class of Scale Mixture of Normal (SMN) distributions is defined by scaling the variance of a normal variable using a positive mixing random variable. A random variable $Y \sim$ SMN(μ,σ^2,ν) admits the following stochastic representation:

$$Y \overset{d}{=} \mu + U^{-1/2}Z$$

where $Z \sim \mathcal{N}(0,\sigma^2)$ and $U \sim H(\cdot;\nu)$ are independent, and $\overset{d}{=}$ denotes equality in distribution. The parameter ν controls the tail behavior of the SMN distribution.
Alternatively, the hierarchical form is given by:

$$Y \mid (U=u) \sim \mathcal{N}(\mu, u^{-1}\sigma^2), \quad U \sim H(u;\nu).$$

Thus, the pdf of the SMN distribution can be expressed as:

$$f_{\text{SMN}}(y;\mu,\sigma^2,\nu) = \int_0^\infty \phi(y;\mu,u^{-1}\sigma^2)\,dH(u;\nu), \quad y \in R.$$

The SMN class provides a flexible tool for robust statistical modeling. It includes several well-known heavy-tailed distributions as special cases, such as the Student's t, Slash, Contaminated Normal, Laplace, Variance-Gamma, and Tail-Inflated Normal distributions.

2.2 Model formulation

Suppose the random errors in the Mixture of Experts (MoE) model follow the Scale Mixture of Normal (SMN) distribution. Then, the probability density function (pdf) of the MoE-SMN model can be written as:

$$f(y_i;\Theta) = \sum_{j=1}^{g} \pi_j(r_i;\tau) f_{\text{SMN}}\left(y_i; x_i^\top \beta_j, \sigma_j^2, \nu_j\right), \quad i = 1,\dots,n,$$

where:

- $\Theta = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_g, \boldsymbol{\tau})$ is the collection of all model parameters,
- $\boldsymbol{\theta}_j = (\boldsymbol{\beta}_j, \sigma_j^2, \nu_j)$ are the parameters of the j-th expert,
- $\pi_j(\mathbf{r}_i; \boldsymbol{\tau})$ is the gating function given by the multinomial logistic link:

$$\pi_j(\mathbf{r}_i; \boldsymbol{\tau}) = \frac{\exp(\boldsymbol{\tau}_j^\top \mathbf{r}_i)}{1 + \sum_{\ell=1}^{g-1} \exp(\boldsymbol{\tau}_\ell^\top \mathbf{r}_i)}, \quad j = 1, \ldots, g-1,$$

  and $\pi_g(\mathbf{r}_i; \boldsymbol{\tau}) = 1 - \sum_{j=1}^{g-1} \pi_j(\mathbf{r}_i; \boldsymbol{\tau})$.

To account for censoring, assume that for each observation we observe (c_i, ρ_i), where:

- c_i is the observed value: either the actual response if uncensored ($y_i = c_i$), or the censoring threshold if $y_i \leq c_i$,
- ρ_i is the censoring indicator: $\rho_i = 0$ if uncensored, and $\rho_i = 1$ if censored.

Then, the likelihood function under left-censoring is given by:

$$L(\Theta \mid \mathbf{y}_{\text{obs}}, \boldsymbol{\rho}) = \prod_{i=1}^{n} \sum_{j=1}^{g} \pi_j(\mathbf{r}_i; \boldsymbol{\tau}) \left[\sigma_j^{-1} f_{\text{SMN}}(e_{ij}; \nu_j)\right]^{1-\rho_i} \left[F_{\text{SMN}}(e_{ij}^c; \nu_j)\right]^{\rho_i},$$

where:

$$e_{ij} = \frac{y_i - \mathbf{x}_i^\top \boldsymbol{\beta}_j}{\sigma_j}, \quad e_{ij}^c = \frac{c_i - \mathbf{x}_i^\top \boldsymbol{\beta}_j}{\sigma_j}.$$

To facilitate Bayesian computation, a hierarchical representation is used:

$$\begin{aligned} Y_i \mid (Z_{ij} = 1, U_i = u_i) &\sim \mathcal{N}(\mathbf{x}_i^\top \boldsymbol{\beta}_j, u_i^{-1} \sigma_j^2), \\ U_i \mid (Z_{ij} = 1) &\sim H(u; \nu_j), \\ \mathbf{Z}_i \mid \mathbf{r}_i &\sim \text{Multinomial}(1; \pi_1(\mathbf{r}_i; \boldsymbol{\tau}), \ldots, \pi_g(\mathbf{r}_i; \boldsymbol{\tau})). \end{aligned}$$

This hierarchical form allows efficient sampling in a Bayesian framework using data augmentation strategies.

2.3 Priors and Hyper-parameter Specifications

In Bayesian analysis, priors incorporate prior knowledge from previous observations, offering important insights. For clarity and computational efficiency, we employ conjugate and improper priors in our MoE-SMN-CR model. Conjugate priors are beneficial as their posteriors remain within the same distribution family, streamlining calculations while preserving interpretability.
Assuming prior independence among all parameters $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_g, \boldsymbol{\tau}\}$, the joint prior distribution factorizes as:

$$\pi(\Theta) = \prod_{j=1}^{g} \pi(\boldsymbol{\beta}_j)\, \pi(\sigma_j^2)\, \pi(\nu_j) \times \prod_{j=1}^{g-1} \pi(\boldsymbol{\tau}_j).$$

The specific prior distributions are as follows:

- **Regression Coefficients:** For each expert $j = 1, \ldots, g$, the regression coefficients $\boldsymbol{\beta}_j$ follow a multivariate normal prior:

$$\boldsymbol{\beta}_j \sim \mathcal{N}_p(\mathbf{0}, \kappa_\beta^2 \mathbf{I}_p),$$

  where κ_β^2 is a large constant (e.g., 100 or 1000) and $\mathbf{I}_p$ is the $p \times p$ identity matrix.

- **Gate Function Parameters:** To enable efficient inference for the multinomial logistic regression governing the gating network, we employ the Pólya-Gamma (PG) data augmentation technique proposed by Polson et al. (2013). Specifically, we assign:

$$\boldsymbol{\tau}_j \sim \mathcal{N}_q(\mathbf{0}, 2000\, \mathbf{I}_q), \quad j = 1, \ldots, g-1,$$

  where $\boldsymbol{\tau}_j$ are the gating coefficients and $\mathbf{I}_q$ denotes the $q \times q$ identity matrix. These priors are weakly informative and facilitate posterior inference via the PG augmentation framework.

- **Pólya-Gamma Data Augmentation:** A random variable $\omega \sim \mathrm{PG}(\xi_1, \xi_2)$ admits the following stochastic representation:

$$\omega \overset{d}{=} \frac{1}{2\pi^2} \sum_{k=1}^{\infty} \frac{g_k}{(k-0.5)^2 + \xi_2^2/(4\pi^2)},$$

where $g_k \sim \mathrm{Gamma}(\xi_1, 1)$ are i.i.d. gamma variables. For a PG-distributed ω, the binomial likelihood involving a log-odds term ψ satisfies the identity:

$$\frac{e^{a\psi}}{(1+e^{\psi})^{\xi_1}} = 2^{-\xi_1} e^{\kappa\psi} \int_0^{\infty} e^{-\omega\psi^2/2} f(\omega; \xi_1)\, d\omega, \quad \text{with } \kappa = a - \xi_1/2.$$

This representation transforms the binomial likelihood into a conditionally Gaussian form, enabling efficient Gibbs updates.

- **Ultimate Pólya-Gamma Sampling:** To further enhance sampling efficiency and reduce posterior autocorrelation, we utilize the Ultimate Pólya-Gamma (UPG) sampler . This extension optimizes latent variable sampling and improves mixing in hierarchical mixture models such as MoE-SMN-CR.

2.4 Full Conditional Posteriors

Due to the complex likelihood structure, the marginal posterior distributions of the parameters are analytically intractable. MCMC methods, such as Gibbs sampling and the Metropolis-Hastings (MH) algorithm, can be employed to generate posterior samples from the full conditional distributions. The full conditional distribution of each parameter, including latent variables, given the value of all remaining ones, is required. Below, we outline how the MCMC algorithm, coupled with a data-augmentation scheme, is used to obtain the full conditional distribution of each parameter in the model:

- **Step 1:** Sample the mixing random variable U_i conditioning on all parameters and latent variables in the model.
  - The full conditional distribution of U_i given $(y_i, Z_i, \dots)$ is determined through the distribution of the mixing random variable U.
- **Step 2:** Treat m censored values as missing data.
  - For a left-censored observation c_k, the conditional distribution of Y_k is:

$$Y_k \mid (y_k < c_k, Z_k = 1, u_k, \beta_j, \sigma_j^2) \sim \mathrm{TN}(x_k\beta_j, u_k^{-1}\sigma_j^2; (-\infty, c_k]),$$

    where $k = 1, \dots, m$.
  - For right-censored observations, the truncation interval switches to $[c_k, +\infty)$.
- **Step 3:** Simulate samples of β_j for $j = 1, \dots, g$ from the posterior distribution:

$$\beta_j \mid (y_{\text{obs}}, Z, \mathbf{u}, \Theta_{(-j)}, \sigma_j^2) \sim \mathcal{N}_p \left(\tilde{\Sigma}_{\beta_j} \left[\sigma_j^{-2} \left(\sum_{i=1}^{n} z_{ij} u_i y_i x_i^\top \right) + \Sigma_{\beta_j}^{-1} b_{\beta_j} \right], \tilde{\Sigma}_{\beta_j} \right),$$

where $\Theta_{(-j)}$ denotes all parameters excluding β_j, and

$$\tilde{\Sigma}_{\beta_j} = \left(\sigma_j^{-2} \sum_{i=1}^{n} z_{ij} u_i x_i x_i^\top + \Sigma_{\beta_j}^{-1} \right)^{-1}.$$

- **Step 4:** Draw samples of σ_j^2 for $j = 1, \dots, g$ from the posterior distribution:

$$\sigma_j^2 \mid (y_{\text{obs}}, Z, u, \Theta_{(-j)}, \beta_j) \sim \mathrm{IG} \left(a_{\sigma_j} + \sum_{i=1}^{n} z_{ij}, \quad b_{\sigma_j} + \sum_{i=1}^{n} z_{ij} u_i (y_i - x_i^\top \beta_j)^2 \right),$$

where $\mathrm{IG}(\cdot, \cdot)$ denotes the inverse-gamma distribution, and $\Theta_{(-j)}$ represents all parameters excluding the j-th component.

- **Step 5:** Sample ν_j based on the assumed mixing distribution $H(u; \nu)$ and its prior specification.
  - The procedure for sampling ν_j depends on the specific form of the mixing distribution. Implementation details for different cases are given in Section 2.5 of the main text.
- **Step 6:** Draw samples of the latent cluster indicator Z_i from its posterior distribution:

  $$Z_i \mid (y_i, r_i, \cdots) \sim \mathcal{M}(1; p_{i1}^{\text{Bayes}}, \ldots, p_{ig}^{\text{Bayes}}),$$

  where $\mathcal{M}$ denotes the multinomial distribution with one trial, and

  $$p_{ij}^{\text{Bayes}} = \frac{\pi_j(r_i; \tau_j)\, f_{\text{SMN}}(y_i; x_i^\top \beta_j, \sigma_j^2, \nu_j)}{\sum_{l=1}^{g} \pi_l(r_i; \tau_l)\, f_{\text{SMN}}(y_i; x_i^\top \beta_l, \sigma_l^2, \nu_l)},$$

  where f_{SMN} is the density function of the skewed multivariate normal distribution.
  - For censored observations, Step 2 is used to impute values for y_i.
  - If $Z_{ij} = 1$, the i-th observation is assigned to the j-th component.
- **EM-based Posterior Membership Probabilities:**
  When using the EM algorithm, the posterior membership probability of the i-th sample belonging to group j is given by:

  $$p_{ij}^{(EM)} = \frac{\pi_j(r_k; \tau_j) f_{\text{SMN}}(\epsilon_{i,j}^c)}{\sum_{l=1}^{g} \pi_l(r_k; \tau_l) f_{\text{SMN}}(\epsilon_{i,l}^c)},$$

  for $k = 1, \ldots, m_d$, where $\epsilon_{i,j}^c$ is the latent residual corresponding to the censored observation, as defined in Equation (4) of the original text. This expression involves evaluating the cdf of the SMN distribution at the censoring threshold.
- **Step 7:** Sample τ using the UPG (Uncollapsed Polya-Gamma) sampler for multinomial data.
  - The g-th group is used as the baseline category.
  - Sampling follows "Algorithm 2" in Zens et al. (2023), employing a three-level hierarchical update scheme, described in Appendix E of the supplementary material.

These full conditionals allow efficient implementation of a Gibbs sampler (with Metropolis-Hastings steps as needed) for posterior inference under the proposed Bayesian MoE-SMN-CR model.

2.5 Sampling Strategy for the Indexing Parameter

In the MoE-SMN-CR model, each expert (component) is associated with a specific error distribution from the Scale Mixture of Normals (SMN) family. These distributions are parameterized by an **indexing parameter** ν_j, which determines the heaviness of the tail — i.e., how robust each expert is to extreme observations. The parameter ν_j is critical in adjusting the behavior of the distribution to handle outliers effectively.

Since ν_j affects the mixing distribution in the SMN formulation, its posterior distribution often lacks a closed form, requiring specialized sampling methods. Depending on the distributional form (Slash, Contaminated Normal, Student's t, Variance-Gamma, or Tail-Inflated Normal), different strategies are adopted for posterior inference. Some distributions (like Slash or CN) allow conjugate updates, while others require Metropolis-Hastings (MH) algorithms. In what follows, we describe the approach for each distribution in detail.

2.5.1 Slash (SL) Distribution

The Slash distribution is a simple heavy-tailed distribution that arises from dividing a standard normal variable by a uniform variable. When a conjugate prior $\nu_j \sim \text{Gamma}(a_{\nu_j}, b_{\nu_j})$ is assigned, the posterior conditional distribution of ν_j is also a Gamma, allowing for direct Gibbs sampling.

$$\nu_j \mid \cdot \sim \text{Gamma}\left(a_{\nu_j} + \sum_i z_{ij},\ b_{\nu_j} - \sum_i z_{ij} \log u_i\right)$$

This closed-form update is computationally convenient and ensures rapid convergence during MCMC sampling.

2.5.2 Contaminated Normal (CN) Distribution

The CN model introduces robustness by mixing two normal distributions — one for regular observations and one with inflated variance to accommodate outliers. Here, ν_j represents the mixing proportion and is modeled using a Beta prior, while an auxiliary variable γ_j determines the contamination threshold and is drawn from a Uniform(0,1) prior.

$$\nu_j \sim \text{Beta}(a_{\nu_j}, b_{\nu_j}), \quad \gamma_j \sim \text{Uniform}(0,1)$$

The full conditionals are:

$$\nu_j \mid \cdot \sim \text{Beta}\left(a_{\nu_j} + \sum_i \frac{z_{ij}(1-u_i)}{1-\gamma_j},\; b_{\nu_j} + \sum_i \frac{z_{ij}(u_i - \gamma_j)}{1-\gamma_j}\right)$$

$$\gamma_j \mid \cdot \sim \text{TG}\left(\frac{1}{2}\sum_i b_{ij} z_{ij} + 1,\; \frac{1}{2}\sum_i b_{ij} z_{ij} e_{ij}^2;\; (0,1)\right)$$

where $b_{ij} = 1$ if $u_i \leq \gamma_j$ and 0 otherwise. This hierarchical setup allows different experts to handle varying contamination levels, improving clustering robustness.

2.5.3 Student's t Distribution

The Student's t distribution is widely used for its ability to model heavy tails with a single parameter ν_j. However, no closed-form exists for the full conditional of ν_j. Thus, two distinct strategies are proposed: the **Hierarchical** (Hier) method and the **Non-Hierarchical** (non-Hier) method.

Hierarchical Method:
This method relies solely on the latent scale variables u_i, making it computationally efficient, but less informative as it does not use the observed response y_i. The log-posterior is expressed as:

$$\pi(\nu_j \mid \cdot) \propto \left(\frac{\nu_j}{2}\right)^{\sum_i z_{ij}} \Gamma\left(\frac{\nu_j}{2}\right)^{-\sum_i z_{ij}} \prod_i u_i^{z_{ij}(\nu_j/2-1)} \exp\left(-\frac{\nu_j}{2}\sum_i z_{ij} u_i\right) \pi(\nu_j)$$

We construct a second-order Taylor expansion:

$$q(\nu_j) = \frac{\nu_j}{2}\sum_i z_{ij} \log\left(\frac{\nu_j}{2}\right) - \sum_i z_{ij} \log\Gamma\left(\frac{\nu_j}{2}\right) - \frac{\nu_j}{2}\sum_i z_{ij}\left(u_i - \log u_i\right) + \log\pi(\nu_j)$$

From $q(\nu_j)$, a truncated normal proposal is defined:

$$\nu_j^{\text{new}} \sim \text{TN}(\mu_\nu, \sigma_\nu^2; (2,40)), \quad \mu_\nu = \nu_j - \frac{q'(\nu_j)}{q''(\nu_j)}, \quad \sigma_\nu^2 = -\frac{1}{q''(\nu_j)}$$

The MH acceptance probability becomes:

$$\alpha = \min\left(1,\; \frac{\exp(q(\nu_j^{\text{new}})) \cdot p(\nu_j)}{\exp(q(\nu_j)) \cdot p(\nu_j^{\text{new}})}\right)$$

Non-Hierarchical Method:
This method improves robustness by incorporating full data likelihood (including y_i, censoring, and gating functions). The posterior becomes:

$$q(\nu_j) = \log(\pi(\nu_j)) + \sum_i \log\left[\pi_j(r_i;\tau) f_T(y_i; x_i^\top \beta_j, \sigma_j^2, \nu_j)^{1-\rho_i} F_T(c_i; x_i^\top \beta_j, \sigma_j^2, \nu_j)^{\rho_i}\right]$$

Several priors for ν_j can be used:

- **Jeffreys prior:**

$$\pi(\nu_j) \propto \left(\frac{\nu_j}{\nu_j+3}\right)^{1/2}\left[\psi'\left(\frac{\nu_j}{2}\right) - \psi'\left(\frac{\nu_j+1}{2}\right)\right] - \frac{2(\nu_j+3)}{\nu_j(\nu_j+1)^2}$$

- **Truncated Gamma:** $\nu_j \sim \text{TG}(a_{\nu_j}, b_{\nu_j};\ (2,40])$

- **Pareto:** $\pi(\nu_j) \propto \frac{2}{\nu_j^2},\ \nu_j \geq 2$

- **Juárez and Steel Prior:**

$$\pi(\nu_j) \propto \frac{\nu_j - 1}{\left(\nu_j - 1 + \frac{4}{1+\sqrt{2}}\right)^3}, \quad \nu_j \geq 1$$

This method is more informative and leads to better inference, particularly in heterogeneous datasets.

2.5.4 Variance-Gamma (VG) Distribution

This distribution supports similar hierarchical and non-hierarchical methods.

- **Hierarchical Approach:** For the Variance-Gamma (VG) distribution, the indexing parameter ν_j can be sampled using hierarchical (Hier) and non-hierarchical (non-Hier) methods, similar to the approach used for the Student's t distribution.

  In the Hierarchical approach, the full conditional distribution is proportional to:

$$\pi(\nu_j \mid Z, u, \text{others}) \propto \left\{ \left(\frac{\nu_j}{2}\right)^{\frac{\nu_j}{2}} \frac{1}{\Gamma\left(\frac{\nu_j}{2}\right)} \right\}^{\sum_{i=1}^{n} z_{ij}} \cdot \nu_j^{a_{\nu_j}-1} \cdot \exp\left(-\frac{\nu_j}{2b_{\nu_j}}\right) \cdot I_{(0,40)}(\nu_j),$$

  where $I_{(0,40)}(\cdot)$ is the indicator function ensuring $\nu_j \in (0, 40)$.

- **Non-Hierarchical Approach:** For the non-Hierarchical case, the conditional posterior distribution becomes:

$$\pi(\nu_j \mid y_{\text{obs}}, \rho, \text{others}) \propto \pi(\nu_j) \prod_{i=1}^{n} \left[f_{\text{VG}}(y_i; x_i\beta_j, \sigma_j^2, \nu_j)^{z_{ij}(1-\rho_i)} \cdot \left(\frac{\pi_j(r_i; \tau)}{F_{\text{VG}}(c_i; x_i\beta_j, \sigma_j^2, \nu_j)} \right)^{z_{ij}\rho_i} \right],$$

  where:

  - $f_{\text{VG}}(\cdot)$ is the density function of the VG distribution,
  - $F_{\text{VG}}(\cdot)$ is its cumulative distribution function (cdf),
  - $\rho_i = 1$ if y_i is censored and 0 otherwise,
  - c_i is the censoring threshold for y_i.

2.5.5 Tail-Inflated Normal (TIN) Distribution

This recently proposed distribution inflates the tails of a normal distribution using a random threshold. Assuming a prior $\nu_j \sim \text{Uniform}(0, 1)$, the full conditional becomes:

$$\pi(\nu_j \mid \cdot) \propto \left(\frac{1}{\nu_j}\right)^{\sum_i z_{ij}} I_{(1-\min_{z_{ij}=1} u_i,\ 1)}(\nu_j)$$

Since this does not belong to a known family, sampling is performed using MH.

2.5.6 Laplace and Normal Distributions

Both of these are special cases in the SMN class where the tail behavior is fixed. As they do not involve an indexing parameter ν_j, no sampling is needed for these components.

Conjugate updates are available in limited cases, but more commonly, MH-based methods are employed. Hierarchical modeling allows each expert to adaptively learn its robustness from the data, enhancing interpretability and performance in real-world applications where outliers and non-Gaussian noise are prevalent.

3 Simulations

3.1 Comparison of MH Strategies for Sampling Indexing Parameters in MoE-T-CR and MoE-VG-CR Models

3.1.1 Objective

The primary objective of this simulation study is to compare the effect of the proposed Metropolis-Hastings (MH) strategies for sampling indexing parameters in the MoE-T-CR (Mixture of Experts with T-distribution Censored Responses) and MoE-VG-CR (Mixture of Experts with Variance-Gamma Censored Responses) models. The focus is on evaluating the variation of the estimated indexing parameters under two different scenarios:

- **Equal scenario** where the parameters $\nu_1 = \nu_2 = \nu_3 = 5$.
- **Unequal scenario** where the parameters $(\nu_1, \nu_2, \nu_3) = (3, 6, 10)$.

3.1.2 Methodology

To evaluate the performance of the proposed MH strategies, a total of 100 Monte Carlo (MC) samples of size 500 were generated from the three-component MoE-T-CR and MoE-VG-CR models. The following parameter values were specified for the simulation:

$$\begin{aligned}
&\beta_1 = (0, 4), \quad \beta_2 = (0, -3), \quad \beta_3 = (-2, 1), \quad (\sigma_1^2, \sigma_2^2, \sigma_3^2) = (0.2, 0.1, 0.2)\\
&\tau_1 = (-2, 4), \quad \tau_2 = (-2, -4)
\end{aligned}$$

The covariates were generated as $r_i = (1, x_{1i})$, where x_{1i} was drawn from a uniform distribution $U(-2, 2)$. For each synthetic dataset, a left-censoring scheme with a level of 15% was implemented, following the generation algorithm described in Mirfarah et al. (2021).

3.1.3 MCMC Sampling

The MCMC sampling procedure was conducted to obtain Bayesian estimates of the indexing parameters. Convergence was typically achieved after collecting 20,000 MCMC samples, with the initial 5,000 iterations discarded as burn-in. The posterior mean of the indexing parameters was computed for each converged MCMC sample. The experiment was repeated over 100 replications to obtain reliable estimates. The Bayesian estimates were used to compute various accuracy metrics, which are discussed below.

3.1.4 Accuracy Metrics

The performance of the estimation method was assessed using the following metrics:

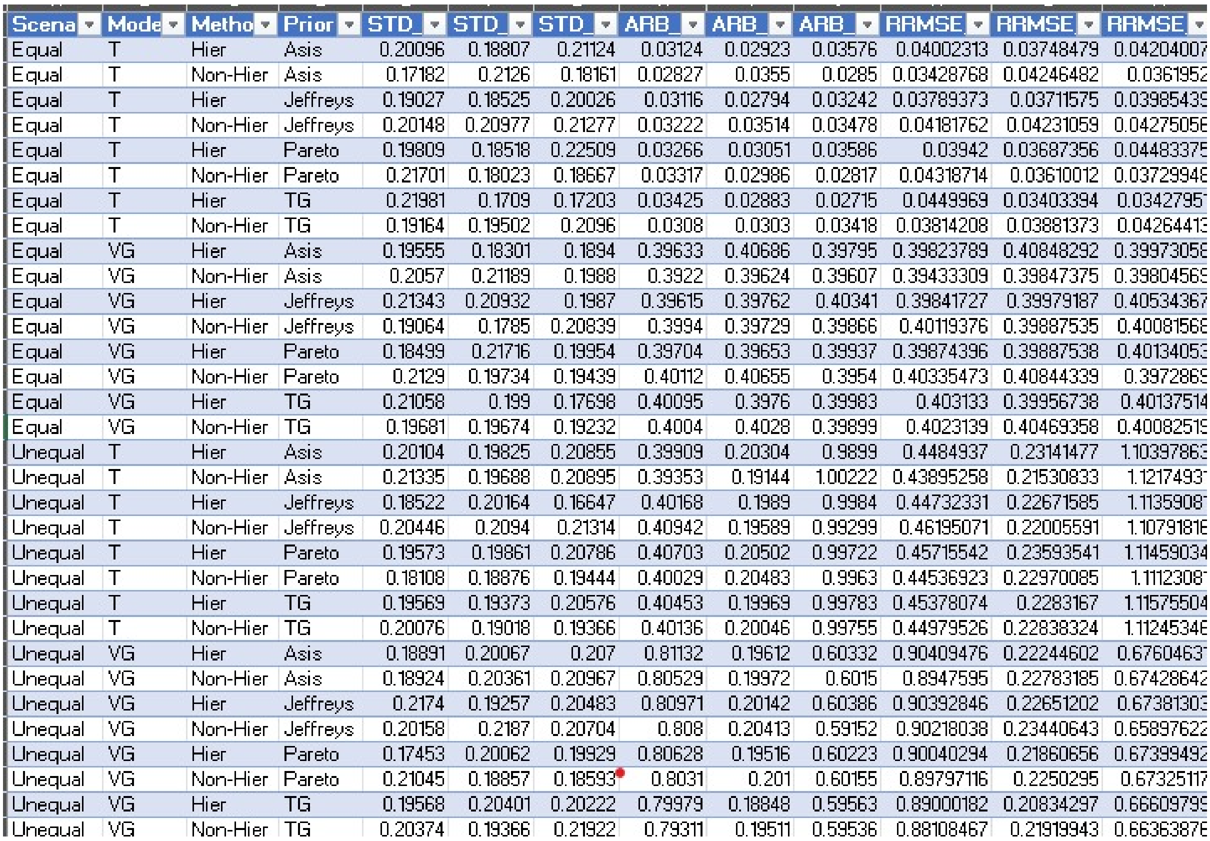
- **Standard Deviation (STD):** The variability of the estimated indexing parameters.
- **Absolute Relative Bias (ARB):** The absolute difference between the true value of ν_j and its Bayesian estimate, normalized by the true value. It is defined as:

$$\text{ARB}_j = \frac{1}{100} \sum_{k=1}^{100} \left| \frac{\hat{\nu_j}^{(k)} - \nu_j}{\nu_j} \right|$$

Where $\hat{\nu_j}^{(k)}$ is the Bayesian estimate (posterior mean) of ν_j obtained at the k^{th} sample.

- **Root Relative Mean Squared Error (RRMSE):** The root mean square error of the Bayesian estimate, normalized by the true value. It is defined as:

$$\text{RRMSE}_j = \sqrt{\frac{1}{100} \sum_{k=1}^{100} \left(\frac{\hat{\nu_j}^{(k)} - \nu_j}{\nu_j} \right)^2}$$



Scena ▾	Mode ▾	Metho ▾	Prior ▾	STD_ ▾	STD_ ▾	STD_ ▾	ARB_ ▾	ARB_ ▾	ARB_ ▾	RRMSE ▾	RRMSE ▾	RRMSE ▾
Equal	T	Hier	Asis	0.20096	0.18807	0.21124	0.03124	0.02923	0.03576	0.04002313	0.03748479	0.04204007
Equal	T	Non-Hier	Asis	0.17182	0.2126	0.18161	0.02827	0.0355	0.0285	0.03428768	0.04246482	0.0361952
Equal	T	Hier	Jeffreys	0.19027	0.18525	0.20026	0.03116	0.02794	0.03242	0.03789373	0.03711575	0.03985439
Equal	T	Non-Hier	Jeffreys	0.20148	0.20977	0.21277	0.03222	0.03514	0.03478	0.04181762	0.04231059	0.04275056
Equal	T	Hier	Pareto	0.19809	0.18518	0.22509	0.03266	0.03051	0.03586	0.03942	0.03687356	0.04483375
Equal	T	Non-Hier	Pareto	0.21701	0.18023	0.18667	0.03317	0.02986	0.02817	0.04318714	0.03610012	0.03729948
Equal	T	Hier	TG	0.21981	0.1709	0.17203	0.03425	0.02883	0.02715	0.0449969	0.03403394	0.03427957
Equal	T	Non-Hier	TG	0.19164	0.19502	0.2096	0.0308	0.0303	0.03418	0.03814208	0.03881373	0.04264413
Equal	VG	Hier	Asis	0.19555	0.18301	0.1894	0.39633	0.40686	0.39795	0.39823789	0.40848292	0.39973058
Equal	VG	Non-Hier	Asis	0.2057	0.21189	0.1988	0.3922	0.39624	0.39607	0.39433309	0.39847375	0.39804569
Equal	VG	Hier	Jeffreys	0.21343	0.20932	0.1987	0.39615	0.39762	0.40341	0.39841727	0.39979187	0.40534367
Equal	VG	Non-Hier	Jeffreys	0.19064	0.1785	0.20839	0.3994	0.39729	0.39866	0.40119376	0.39887535	0.40081568
Equal	VG	Hier	Pareto	0.18499	0.21716	0.19954	0.39704	0.39653	0.39937	0.39874396	0.39887538	0.40134053
Equal	VG	Non-Hier	Pareto	0.2129	0.19734	0.19439	0.40112	0.40655	0.3954	0.40335473	0.40844339	0.3972869
Equal	VG	Hier	TG	0.21058	0.199	0.17698	0.40095	0.3976	0.39983	0.403133	0.39956738	0.40137514
Equal	VG	Non-Hier	TG	0.19681	0.19674	0.19232	0.4004	0.4028	0.39899	0.4023139	0.40469358	0.40082519
Unequal	T	Hier	Asis	0.20104	0.19825	0.20855	0.39909	0.20304	0.9899	0.4484937	0.23141477	1.10397863
Unequal	T	Non-Hier	Asis	0.21335	0.19688	0.20895	0.39353	0.19144	1.00222	0.43895258	0.21530833	1.1217493
Unequal	T	Hier	Jeffreys	0.18522	0.20164	0.16647	0.40168	0.1989	0.9984	0.44732331	0.22671585	1.1135908
Unequal	T	Non-Hier	Jeffreys	0.20446	0.2094	0.21314	0.40942	0.19589	0.99299	0.46195071	0.22005591	1.10791816
Unequal	T	Hier	Pareto	0.19573	0.19861	0.20786	0.40703	0.20502	0.99722	0.45715542	0.23593541	1.11459034
Unequal	T	Non-Hier	Pareto	0.18108	0.18876	0.19444	0.40029	0.20483	0.9963	0.44536923	0.22970085	1.1112308
Unequal	T	Hier	TG	0.19569	0.19373	0.20576	0.40453	0.19969	0.99783	0.45378074	0.2283167	1.11575504
Unequal	T	Non-Hier	TG	0.20076	0.19018	0.19366	0.40136	0.20046	0.99755	0.44979526	0.22838324	1.11245346
Unequal	VG	Hier	Asis	0.18891	0.20067	0.207	0.81132	0.19612	0.60332	0.90409476	0.22244602	0.6760463
Unequal	VG	Non-Hier	Asis	0.18924	0.20361	0.20967	0.80529	0.19972	0.6015	0.8947595	0.22783185	0.67428642
Unequal	VG	Hier	Jeffreys	0.2174	0.19257	0.20483	0.80971	0.20142	0.60386	0.90392846	0.22651202	0.67381303
Unequal	VG	Non-Hier	Jeffreys	0.20158	0.2187	0.20704	0.808	0.20413	0.59152	0.90218038	0.23440643	0.65897622
Unequal	VG	Hier	Pareto	0.17453	0.20062	0.19929	0.80628	0.19516	0.60223	0.90040294	0.21860656	0.67399492
Unequal	VG	Non-Hier	Pareto	0.21045	0.18857	0.18593	0.8031	0.201	0.60155	0.89797116	0.2250295	0.67325117
Unequal	VG	Hier	TG	0.19568	0.20401	0.20222	0.79979	0.18848	0.59563	0.89000182	0.20834297	0.66609799
Unequal	VG	Non-Hier	TG	0.20374	0.19366	0.21922	0.79311	0.19511	0.59536	0.88108467	0.21919943	0.66363876

Figure 1: Simulation results for assessing sampling strategies for the indexing parameter in the MoE-T-CR and MoE-VG-CR models for “equal” and “unequal” scenarios

3.2 Interpretation

3.2.1 Equal Variance Scenario

T-Distribution Models

- The T-distribution models demonstrate superior performance across all prior specifications, with RRMSE values consistently below 0.05 for all parameters (v1, v2, v3).
- Hierarchical models with ”TG” priors yield the lowest RRMSE for parameters v2 (0.034) and v3 (0.034), suggesting optimal performance for these parameters.
- Non-hierarchical models with ”Asis” priors perform best for parameter v1 (RRMSE = 0.034), indicating that different model specifications may be optimal for different parameters.
- The standard deviations (STD) across all T-distribution models remain relatively stable (ranging from 0.17 to 0.22), indicating consistent precision regardless of prior choice.

VG-Distribution Models

- VG-distribution models perform substantially worse than T-distribution models, with RRMSE values approximately 10 times higher (around 0.40) across all parameters.
- The average absolute bias (ARB) for VG models is dramatically higher (approximately 0.39–0.40) compared to T models (approximately 0.03–0.04).
- This substantial performance gap suggests that VG-distribution models are inappropriate for data with equal variance characteristics.

3.2.2 Unequal Variance Scenario

T-Distribution Models

- Under unequal variance conditions, T-distribution models show parameter-specific performance patterns:
  - For v1: RRMSE values increase to approximately 0.44–0.46
  - For v2: RRMSE values increase moderately to approximately 0.22–0.24
  - For v3: RRMSE values increase dramatically to approximately 1.10–1.12
- The substantially higher RRMSE for v3 indicates that this parameter is particularly sensitive to variance heterogeneity.
- Prior choice becomes more influential under unequal variance, though no single prior consistently outperforms others across all parameters.

VG-Distribution Models

- Interestingly, VG models show improved relative performance under unequal variance compared to equal variance scenarios:
  - For v1: RRMSE values are high (approximately 0.88–0.90)
  - For v2: RRMSE values are comparable to T models (approximately 0.21–0.23)
  - For v3: RRMSE values are lower than T models (approximately 0.66–0.68)
- This suggests that while VG models remain suboptimal overall, they may offer advantages for specific parameters (particularly v3) when variances are unequal.

3.2.3 Hierarchical vs. Non-Hierarchical Modeling

- In the equal variance scenario, hierarchical modeling provides marginal improvements for T-distribution models with certain priors, but the benefits are not consistent across all parameters and prior specifications.
- Under unequal variance conditions, the performance difference between hierarchical and non-hierarchical approaches becomes even less pronounced, suggesting limited practical advantage to hierarchical modeling in this context.

3.2.4 Prior Sensitivity

- For T-distribution models with equal variance, the ”Jeffreys” prior offers the most balanced performance across all parameters.
- For unequal variance scenarios, prior choice has a more substantial impact, though no single prior demonstrates clear superiority.
- The sensitivity to prior specification increases under challenging estimation conditions (unequal variance), highlighting the importance of careful prior selection in complex modeling scenarios.

3.3 Comparison of Bayesian and EM-based Inferences

The Bayesian approach can potentially offer improved clustering performance, especially when dealing with censored response variables. To evaluate this, we conduct a comparative analysis between our proposed Bayesian inference method and the likelihood-based Expectation-Maximization (EM) algorithm.

In this experiment, we generate synthetic datasets of size $n = 500$ from a parallel 3-component Mixture of Experts model with Skewed Mixture Normal-Censored Regression (MoE-SMN-CR) structure. The true parameter values used in the simulation are as follows:

- **Regression coefficients:**
  $\boldsymbol{\beta}_1 = (10, 4), \quad \boldsymbol{\beta}_2 = (0, 4), \quad \boldsymbol{\beta}_3 = (-10, 4)$
- **Gating network coefficients:**
  $\boldsymbol{\tau}_1 = (-2, 4), \quad \boldsymbol{\tau}_2 = (-2, -4)$

- **Error variances:**
  $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = 1$

Covariates are defined as $\boldsymbol{x}_i = (1, x_{1i})$ and $\boldsymbol{r}_i = (1, r_{1i})$, where both x_{1i} and r_{1i} are independently drawn from a uniform distribution $U(-2, 2)$. To ensure a fair comparison between the two inference methods, the mixing variable U_i is simulated from the Generalized Inverse Gaussian (GIG) distribution, as introduced by Good (1953).

The indexing parameters for the three components are:
$\boldsymbol{\nu}_1 = (-0.5, 1, 0.2), \quad \boldsymbol{\nu}_2 = (0.5, 1, 0.2), \quad \boldsymbol{\nu}_3 = (-0.5, 1, 0.2)$

This setup yields samples from a mixture of symmetric geometric hyperbolic distributions, enabling a robust comparison of the clustering performance achieved by the Bayesian and EM-based approaches.

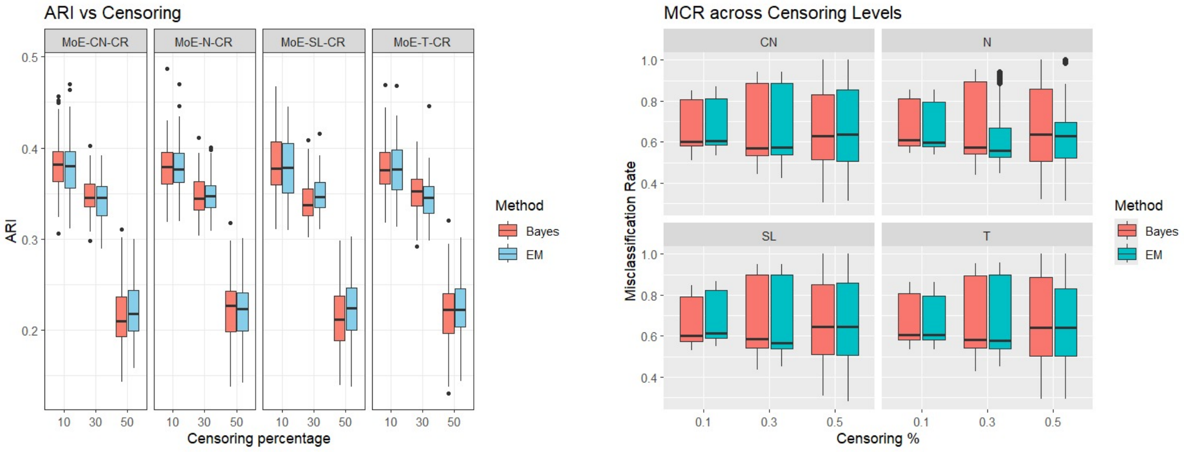


Figure 2: Comparison of the likelihood inference via EM algorithm and Bayes method for four particular cases of the MoE-SMN-CR model in terms of the MCR and ARI scores

3.4 Interpretation

As censoring increases from 10% to 50%, clustering performance deteriorates, evidenced by a steady decline in Adjusted Rand Index (ARI) and a corresponding rise in Misclassification Rate (MCR) across all models. red-Bayesian methods (red) consistently outperform blueEM algorithms (blue), offering higher median ARI and lower MCR with reduced variability. Among the models, heavy-tailed error distributions—specifically `MoE-T-CR`, `MoE-SL-CR`, and `MoE-CN-CR`—demonstrate superior robustness compared to the Gaussian model `MoE-N-CR`. Notably, `MoE-CN-CR` tends to deliver the best performance in both ARI and MCR metrics. Overall, Bayesian MCMC approaches paired with heavy-tailed models provide the most accurate and stable clustering results under varying levels of censoring.

3.5 Sensitivity Analysis of Parameter Estimation in the Presence of Outliers

An additional simulation study was conducted to assess the robustness of the proposed Bayesian estimation procedure under data contamination. Robustness of estimation methods in the presence of outliers has been extensively studied in the literature using approaches such as case-deletion diagnostics, local influence, perturbation analysis, and contamination-based studies. Following this tradition, our objective here is to investigate the stability of parameter estimates when the response variable is subject to outliers.

To perform this analysis, 100 Monte Carlo samples were generated, each consisting of $n = 500$ observations drawn from a 5-component Mixture of Experts with Normal experts and Censored Regression (MoE-N-CR) model. The true parameter values for the simulation are specified as follows:

Expert coefficients:

$$\beta_1 = (6, -2, 3, 2, -3), \quad \beta_2 = (-3, 1, 1, -2, 2), \quad \beta_3 = (3, -1.5, -2, -2, -1),$$
$$\beta_4 = (-0.5, -2, 3, 3, -2), \quad \beta_5 = (-6, -2, -1, -3, 2)$$

Gating parameters:

$$\begin{aligned}\tau_1 &= (1, 0.5, -1.5, 2),\\ \tau_2 &= (2, -0.2, 2, 1),\\ \tau_3 &= (-2, 3, -2, 1),\\ \tau_4 &= (-2, -3, -0.7, -2)\end{aligned}$$

Error variances: $(\sigma_1^2, \sigma_2^2, \sigma_3^2, \sigma_4^2, \sigma_5^2) = (1, 2, 2, 1, 3)$.

The covariates for the experts and gating functions were generated as:

$$x_i = (1, x_{1i}, x_{2i}, x_{3i}, x_{4i}), \quad r_i = (1, r_{1i}, r_{2i}, x_{4i}),$$

where

$$x_{1i} \sim U(-1, 1), \quad x_{2i} \sim \text{Poisson}(2), \quad x_{3i} \sim \mathcal{N}(0, 1), \quad x_{4i} \sim U(-2, 2), \quad r_{1i}, r_{2i} \sim \mathcal{N}(0, 1).$$

To simulate the presence of outliers, a contamination process was applied by randomly selecting $\delta \in \{10, 20, 40, 60\}$ observations from each sample and replacing their corresponding responses with values generated outside the typical range of the original responses. Specifically, for a contaminated response y_i^*, we generated:

$$y_i^{(\text{out})} \sim U(-20, -15)\, I_{(-\infty,0)}(y_i^*) + U(15, 20)\, I_{[0,\infty)}(y_i^*),$$

ensuring the outliers lie significantly outside the data's natural support.

Subsequently, seven variations of the MoE-SMN-CR model were fitted to both the uncontaminated and contaminated datasets. The resulting parameter estimates from the clean and contaminated datasets are denoted by $\hat{\theta}$ and $\hat{\theta}^{(\delta)}$, respectively.

To quantify the sensitivity of parameter estimation to contamination, we computed the Mean Magnitude of Absolute Error (MMAE) for the regression and gating parameters:

$$\text{MMAE}(\beta) = \frac{1}{25}\sum_{j=1}^{5}\sum_{l=1}^{5}\left|\hat{\beta}_{lj}^{(\delta)} - \hat{\beta}_{lj}\right|, \quad \text{MMAE}(\tau) = \frac{1}{16}\sum_{j=1}^{4}\sum_{l=1}^{4}\left|\hat{\tau}_{lj}^{(\delta)} - \hat{\tau}_{lj}\right|.$$

These metrics evaluate the deviation in parameter estimates due to contamination, thereby serving as a robustness measure for the Bayesian estimation method under consideration.

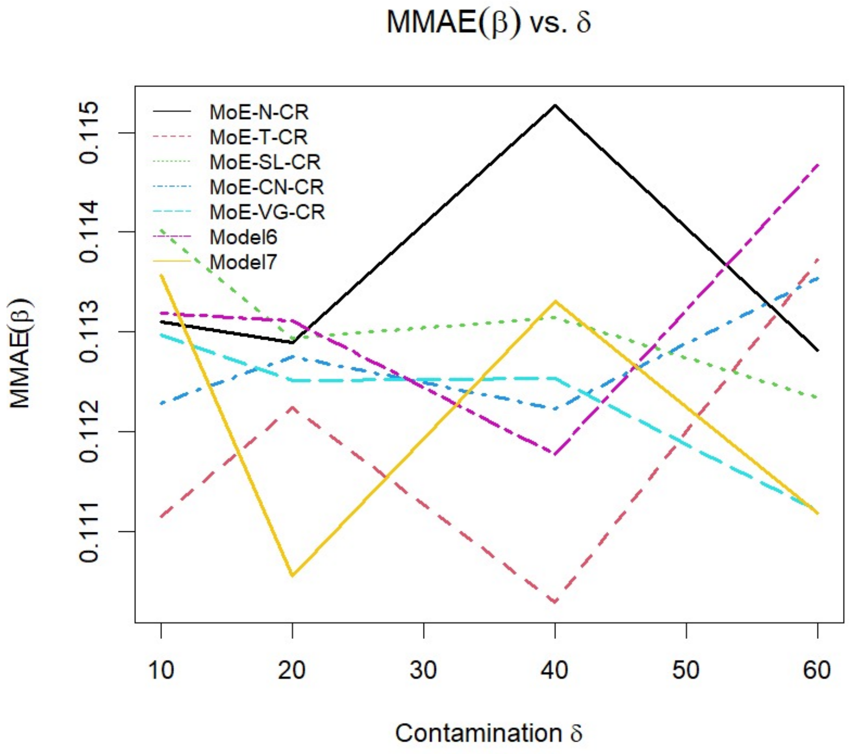


Figure 3: Average MMAE as a function of contamination level (δ) for the covariate β.

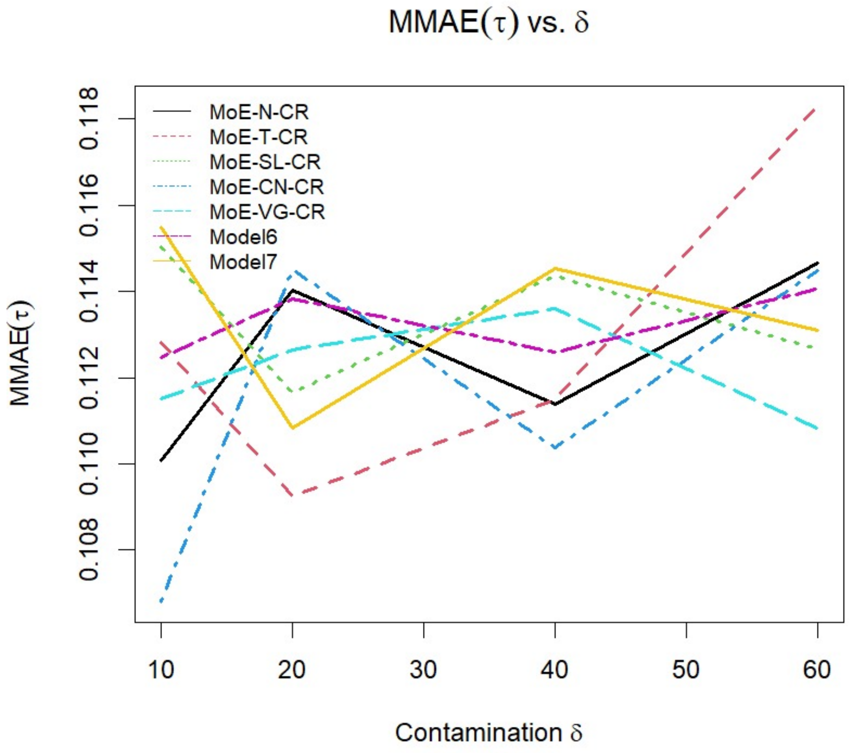


Figure 4: Average MMAE as a function of contamination level (δ) for the gating parameter τ.

3.5.1 Interpretation

- **MMAE(β) vs. δ:**
  - The Gaussian model (MoE-N-CR) exhibits a marked increase in error as contamination grows, indicating low robustness.
  - Heavy-tailed/skewed models (MoE-T-CR, MoE-SL-CR, MoE-CN-CR, MoE-VG-CR) show nearly flat MMAE(β) curves; MoE-SL-CR and MoE-CN-CR are the most stable.
  - Reference models (Model6, Model7) lie between Gaussian and heavy-tailed variants in sensitivity.
- **MMAE(τ) vs. δ:**
  - MoE-N-CR error increases monotonically with outlier level.
  - MoE-CN-CR error slightly *decreases* as contamination rises.
  - Other heavy-tailed models maintain almost constant MMAE(τ), confirming superior robustness.

4 Case study: Wage Data

This case study uses an economic dataset available in an R package that contains wage information for married white women in the U.S., aged 30 to 60, in the year 1975. The key variable of interest is the average hourly wage in dollars. Out of 753 women, 325 (43.16%) reported zero working hours. These are considered left-censored observations, as they reflect negative desired working hours—interpreted as opting out of the labor force.

The response variable y is defined as the annual number of hours the wife worked outside the home, divided by 1000. The explanatory variables include: the wife's age (x_1), log of family income (x_2), regional unemployment rate (x_3), and non-wife income ($x_4 = \frac{\text{family income}-\text{wage}\times\text{hours}}{1000}$). For logistic regression covariates, we set $\mathbf{r}_i^T = (1, x_{2i}, x_{4i})$.

To explore heterogeneity in the data, we applied a finite mixture model using MCMC sampling. Models with different numbers of components ($g = 1$ to 5) were estimated. After discarding the first 10,000 samples as burn-in and retaining 30,000 posterior samples, model selection criteria indicated that the best fit was with $g = 2$. This suggests the presence of two distinct subpopulations in the dataset, which single-component models could not adequately capture.

4.1 Model Interpretation and Comparison

- The table reports the posterior means, standard deviations, and 95% highest posterior density (HPD) intervals for the parameters, as well as model selection criteria for two-component MoE-SMN-CR sub-models.
- The heavy-tailed MoE-SMN-CR models outperform the MoE-N-CR model, which provides the weakest fit due to its limited ability to accommodate outliers.
- The MoE-SL-CR model consistently achieves the best performance across all model selection criteria, demonstrating its robustness in capturing data structure.
- Across all MoE-SMN-CR models, family income shows a positive association with wives' annual work hours in both latent classes, while work hours tend to increase when regional unemployment rate (x_3) and non-wife income (x_4) decrease.
- The gating parameters are moderately significant, indicating that the covariates in the logistic gating function ($\mathbf{r}$) play an important role in distinguishing between the latent subpopulations.
- Among all the models, the T model performed the best across every criterion. It had the highest LPML value of 244.070, indicating superior predictive ability. Furthermore, it achieved the lowest DIC (241.406), EAIC (976.661), EBIC (5847.033), and both WAIC values (-488.130), making it the most parsimonious and accurate model overall. These results confirm that the T model provides the best trade-off between fit and complexity.
- The L model also showed relatively strong performance, with an LPML of 843.353 and moderate DIC (3392.838) and WAIC values (1686.704). While it did not outperform the T model, it was clearly more effective than most of the others, especially in terms of balancing predictive accuracy with model simplicity.

- In contrast, the TIN model demonstrated the worst performance, with the lowest LPML value of 1204.179, indicating poor predictive accuracy. It also had one of the highest DIC (5567.202), EAIC (9058.079), EBIC (32181.872), and WAIC values (2131.672), suggesting overfitting and weak generalization capability.
- Similarly, the Slash model, VG model, and N model all showed inferior performance compared to the T and L models. These models had low LPML values (e.g., 1094.663 for Slash, 1062.693 for VG) and very high DIC and WAIC values, indicating that they were not suitable for accurately modeling the data or capturing its underlying structure.
- The CN model performed slightly better than TIN, Slash, and VG, but still lagged behind the T and L models. Its LPML of 875.131 and DIC of 3456.892 were not competitive enough to recommend its use in practice.

Table 1: Posterior summaries for MoE-N-CR and MoE-L-CR models

	MoE-N-CR Model				MoE-L-CR Model			
Parameter	Mean	SD	HPDlo	HPDhi	Mean	SD	HPDlo	HPDhi
β_{01}	-18.3076	1.7753	-21.8348	-14.8938	-31.4059	1.4853	-34.2898	-28.4770
β_{11}	0.005933	0.004913	-0.00341	0.015891	-0.00517	0.004864	-0.01472	0.004375
β_{21}	2.2083	0.2002	1.8334	2.6189	2.7467	0.1673	2.3482	3.0704
β_{31}	-0.00932	0.012604	-0.0339	0.015509	-0.01213	0.012688	-0.03683	0.01257
β_{41}	-0.1551	0.01772	-0.1902	-0.1203	-0.2839	0.0154	-0.3153	-0.2531
β_{02}	-63.0747	6.2329	-75.4465	-51.7462	-67.8212	5.5941	-79.1164	-57.2496
β_{12}	-0.01571	0.009265	-0.03356	0.00105	-0.005	0.008857	-0.02057	0.01037
β_{22}	6.7101	0.6388	5.5259	7.9701	7.0817	0.5678	5.9073	8.2220
β_{32}	-0.02008	0.02319	-0.06424	0.025791	-0.00654	0.022445	-0.05259	0.03628
β_{42}	-0.1761	0.01661	-0.2083	-0.1429	-0.1694	0.0138	-0.1965	-0.1436
σ_1^2	-76.2150	9.1579	-94.4786	-58.673	42.2764	24.3098	-8.7403	109.7381
σ_2^2	9.2538	1.0728	7.1820	11.3705	1.1303	1.9170	-2.6575	4.9172
τ_{01}	-0.8910	0.0971	-1.0817	-0.7083	-2.4743	0.8646	-4.1651	-1.0456
τ_{11}	0.3466	0.03988	0.2727	0.4274	0.6387	0.0570	0.5313	0.7512
τ_{21}	0.9670	0.1449	0.6916	1.2677	0.7619	0.1109	0.5585	0.9809

Table 2: Posterior summaries for T-CR and SL-CR models

	T-CR Model				SL-CR Model			
Parameter	Mean	SD	HPDlo	HPDhi	Mean	SD	HPDlo	HPDhi
β_{01}	-21.52	3.603	-29.122	-15.669	-19.53	2.55	-24.49	-14.73
β_{11}	0.00274	0.00564	-0.0091	0.01311	0.0043	0.0052	-0.0052	0.0149
β_{21}	2.55465	0.414	1.918	3.465	2.33	0.27	1.82	2.86
β_{31}	-0.0081	0.0128	-0.0362	0.01707	-0.0084	0.0128	-0.0332	0.0168
β_{41}	-0.1627	0.03773	-0.2468	-0.109	-0.1490	0.0187	-0.1857	-0.1130
β_{02}	-41.541	28.4646	-76.584	0.01303	-43.44	25.93	-73.37	-1.82
β_{12}	-0.0129	0.01189	-0.0371	0.01006	-0.0137	0.0117	-0.0374	0.0089
β_{22}	4.36919	3.03798	-0.0548	8.1203	4.59	2.78	0.0403	7.70
β_{32}	-0.0191	0.0301	-0.082	0.0399	-0.0212	0.0296	-0.0828	0.0374
β_{42}	-0.1057	0.08541	-0.2069	0.02726	-0.1147	0.0796	-0.2036	0.0199
σ_1^2	-88.511	26.3201	-133.35	-44.275	-89.99	20.51	-130.79	-60.16
σ_2^2	10.5114	2.644	6.115	15.198	10.62	2.12	7.44	14.93
τ_{01}	-0.853	0.112	-1.0811	-0.6476	-0.8497	0.1027	-1.0558	-0.6500
τ_{11}	0.3939	0.103	0.244	0.603	0.3573	0.0682	0.2487	0.4910
τ_{21}	0.607	1.002	0.5334	3.425	1.5886	1.0347	0.5343	3.5357
ν_1	17.6464	8.2006	6.2449	34.852	10.99	4.70	4.32	19.66
ν_2	23.35	7.80674	9.89272	38.026	13.53	4.00	6.68	20.00

Table 3: Posterior summaries for VG and TIN models

	VG Model				TIN Model			
Parameter	Mean	SD	HPDlo	HPDhi	Mean	SD	HPDlo	HPDhi
β_{01}	-17.82	1.82	-21.65	-14.41	-17.43	1.88	-21.17	-13.73
β_{11}	0.00498	0.00517	-0.00535	0.01481	0.00527	0.00505	-0.00499	0.01478
β_{21}	2.16	0.21	1.76	2.58	2.12	0.21	1.69	2.53
β_{31}	-0.00940	0.01288	-0.03325	0.01729	-0.01150	0.01280	-0.03621	0.01405
β_{41}	-0.14987	0.01805	-0.18629	-0.11528	-0.14874	0.01832	-0.18502	-0.11288
β_{02}	-63.77	6.50	-76.79	-51.90	-60.54	5.87	-71.96	-49.20
β_{12}	-0.01420	0.00929	-0.03252	0.00361	-0.01579	0.00910	-0.03449	0.00151
β_{22}	6.77	0.66	5.56	8.10	6.46	0.60	5.28	7.62
β_{32}	-0.01892	0.02289	-0.06410	0.02649	-0.02124	0.02250	-0.06409	0.02413
β_{42}	-0.17509	0.01675	-0.20909	-0.14398	-0.17191	0.01614	-0.20359	-0.14060
σ_1^2	-80.38	8.77	-97.55	-63.26	-79.40	9.25	-97.92	-61.37
σ_2^2	9.70	1.03	7.72	11.75	9.60	1.08	7.47	11.73
τ_{01}	-0.8988	0.0945	-1.082	-0.716	-0.9073	0.0973	-1.111	-0.728
τ_{11}	0.3702	0.0539	0.277	0.476	0.3469	0.0538	0.2552	0.4694
τ_{21}	0.8907	0.1795	0.573	1.233	0.9713	0.1691	0.6210	1.2860
ν_1	9.86	6.57	2.41	24.11	0.2474	0.1163	0.0509	0.4633
ν_2	14.53	8.41	3.51	32.21	0.2877	0.1160	0.0785	0.5018

Table 4: Posterior summaries for CN model

Parameter	Mean	SD	HPDlo	HPDhi
β_{01}	-29.85	2.10	-33.79	-25.72
β_{11}	-0.00440	0.00509	-0.01409	0.00578
β_{21}	3.56	0.25	3.07	4.02
β_{31}	-0.01108	0.01308	-0.03619	0.01524
β_{41}	-0.26560	0.02746	-0.31571	-0.21431
β_{02}	-66.60	5.62	-77.64	-55.69
β_{12}	-0.00601	0.00873	-0.02328	0.01093
β_{22}	6.98	0.57	5.89	8.11
β_{32}	-0.00876	0.02156	-0.05068	0.03381
β_{42}	-0.17171	0.01440	-0.20016	-0.14411
σ_1^2	0.565	49.62	-65.90	81.59
σ_2^2	3.67	3.31	-1.91	8.97
τ_{01}	-1.8218	1.0104	-3.8768	-0.7138
τ_{11}	0.6151	0.0612	0.5020	0.7391
τ_{21}	0.7434	0.1036	0.5517	0.9447
ν_1	0.0681	0.0499	0.0101	0.1502
ν_2	0.3339	0.1163	0.1170	0.4999
γ_1	0.2493	0.1365	0.0422	0.5073
γ_2	0.8539	0.1264	0.6106	1.0000

Table 5: Model comparison based on LPML, DIC, EAIC, EBIC, WAIC1, and WAIC2

Model	**LPML**	**DIC**	**EAIC**	**EBIC**	**WAIC1**	**WAIC2**
N	-1059.244	5785.983	9514.411	34211.765	2116.579	2116.579
L	-843.353	3392.838	5116.315	16532.737	1686.704	1686.704
T	244.070	241.406	976.661	5847.033	-488.130	-488.130
Slash	-1094.663	6003.077	9944.268	36050.973	2275.359	2275.359
VG	-1062.693	6594.117	11114.469	41057.575	2124.435	2124.435
TIN	-1204.179	5567.202	9058.079	32181.872	2131.672	2131.672
CN	-875.131	3456.892	5223.587	16926.289	1734.556	1734.556

4.2 Plots

4.2.1 ACF plots

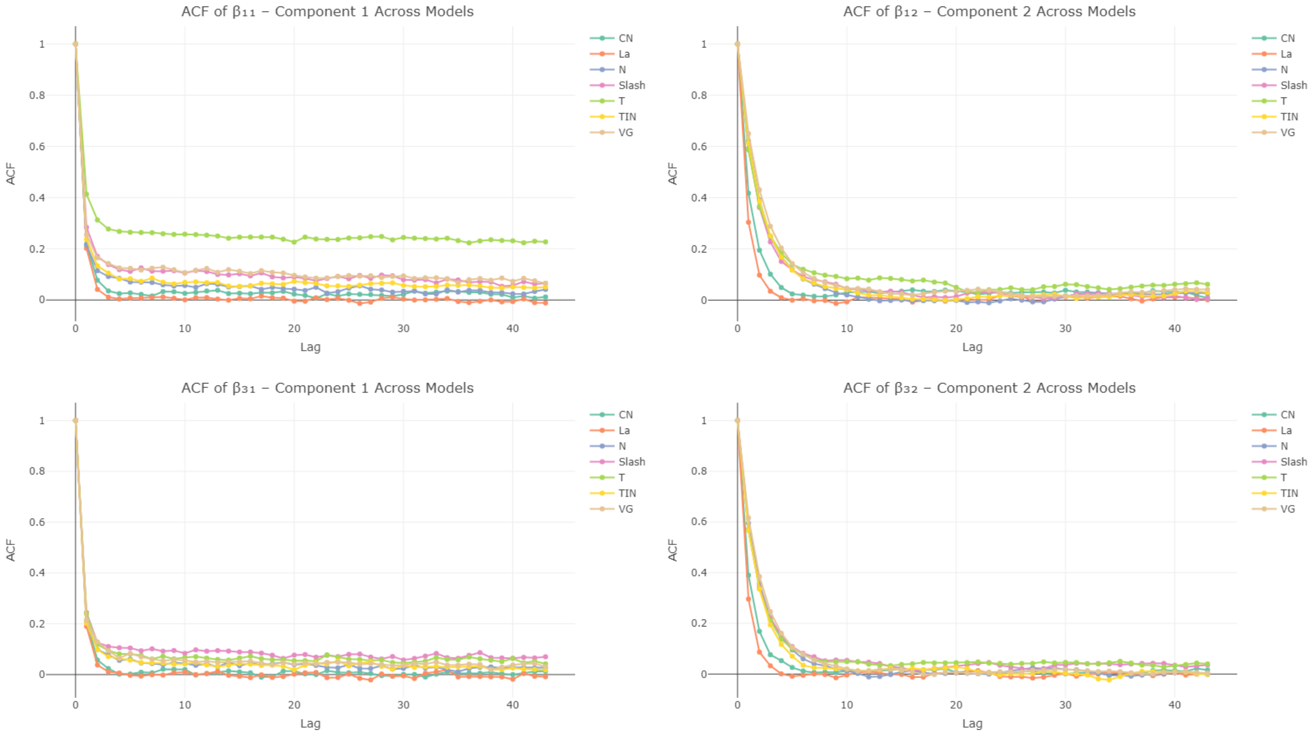


Figure 5: ACF plots for some of the parameters across models

Interpretation :

- The posterior estimates were examined for four key parameters:
  - β_{11}, β_{12}: Effects of the covariate on the **mean** in Components 1 and 2.
  - β_{31}, β_{32}: Effects of the covariate on the **autocorrelation structure (ACF)** in Components 1 and 2.
- Across all error distributions considered (Normal, Student-t, Slash, Contaminated Normal):
  - Estimates of β_{11} and β_{12} are **close to zero**, with **95% credible intervals including zero**,
    $\Rightarrow$ Indicates **no significant covariate effect on the mean response**.
  - Estimates of β_{31} and β_{32} have **posterior means near zero** and **wide credible intervals**,
    $\Rightarrow$ Suggests **minimal influence of covariates on autocorrelation (spatial/temporal dependence)**.
- This trend is **consistent across all mixture model variants**, implying:
  - The components mostly capture **latent structure or background variability**.
  - Rather than being driven by **observed covariate effects**.

4.2.2 Posterior Density Plots

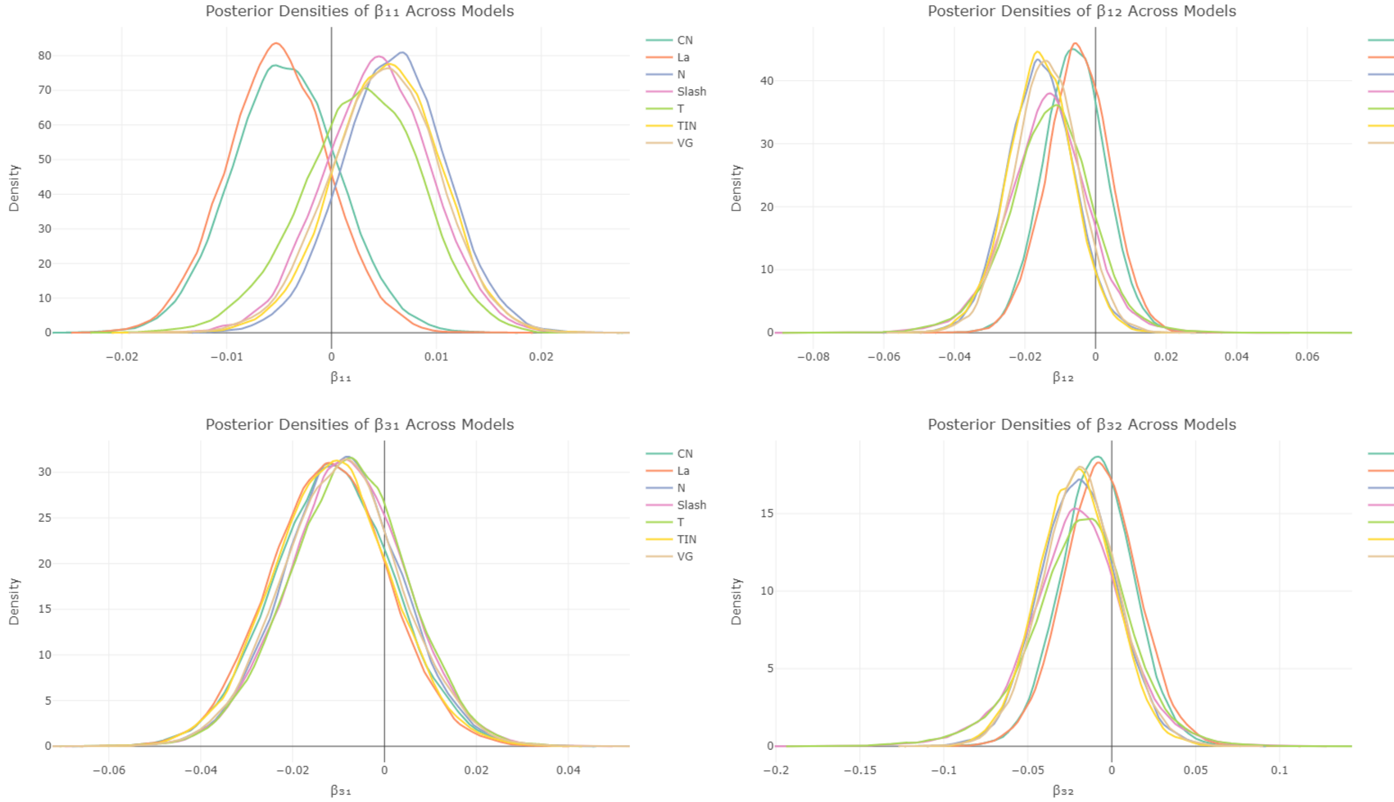
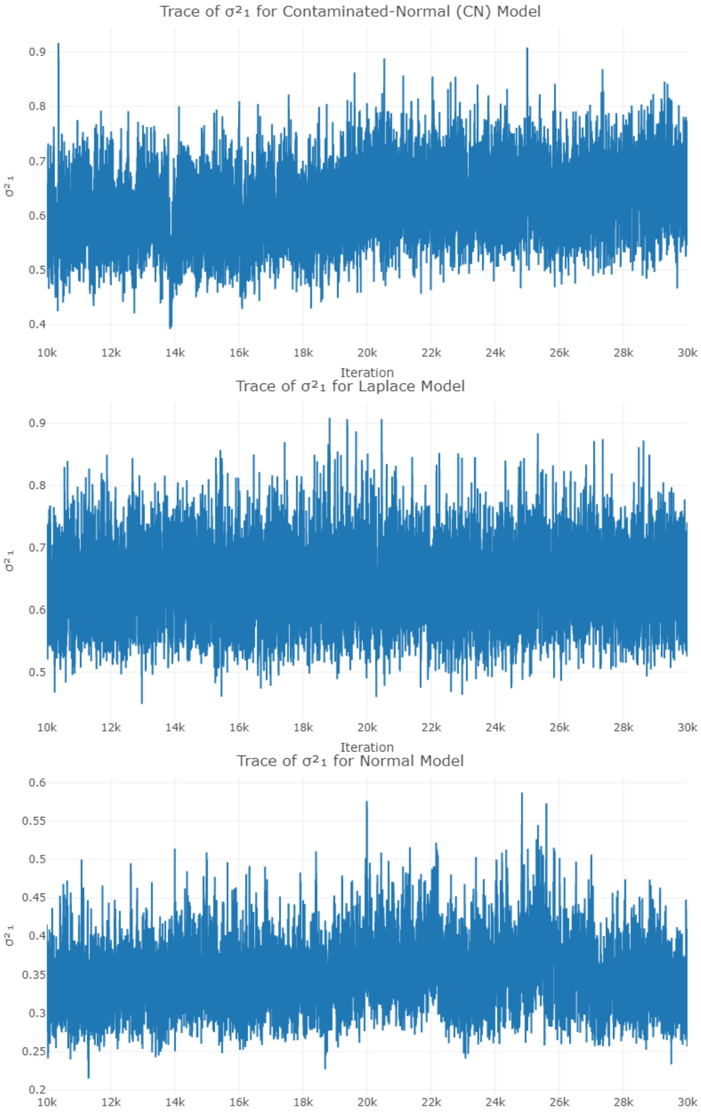


Figure 6: ACF plots for some of the parameters across models

Interpretation:

- **Model Robustness:** The posterior densities of $\beta_{11}, \beta_{12}, \beta_{21}, \beta_{22}$ across all models (CN, La, Slash, T, TIN, VG) are largely overlapping, indicating that parameter estimates are robust to the choice of error distribution.
- **Centered Around Zero:** Most of the densities are sharply peaked around zero, suggesting that the corresponding covariates have minimal or no significant effect on the response variable in both mixture components.
- **Tail Behavior:** Heavier-tailed distributions like Slash, T, and VG show slightly wider posterior spreads, reflecting greater uncertainty, but the central tendency remains consistent across models.
- **Inference Consistency:** The high overlap among models implies that the underlying inference drawn from the regression parameters remains consistent, supporting model reliability regardless of error assumptions.

4.2.3 Trace Plots


Trace of $\sigma^2{}_1$ for Contaminated-Normal (CN) Model

Trace of $\sigma^2{}_1$ for Laplace Model

Trace of $\sigma^2{}_1$ for Normal Model

Figure 7: Trace plots for CN, Laplace and Normal Models

Interpretation:

The trace plots of σ_1^2 under the **Contaminated Normal**, **Laplace**, and **Normal** models exhibit desirable MCMC behavior. All three show good mixing and stationarity, with the Markov chains fluctuating around a stable mean without evident trends or drifts. The Normal model trace is particularly tight and centered, indicating strong convergence with low autocorrelation. The Laplace and Contaminated Normal models also display consistent sampling, suggesting that the posterior distributions for σ_1^2 have been adequately explored under these models. These results indicate reliable and stable parameter estimation for these three error distributions.

5 Conclusion

This paper introduces a fully Bayesian approach to the Mixture of Experts (MoE) model using the Scale Mixture of Normal (SMN) distributions, with the capability to handle left and right censoring of the response variable. By leveraging conjugate and weakly informative priors, we have derived explicit full conditional posterior distributions for parameter estimation. The use of the Ultimate Pólya-Gamma (UPG) data-augmentation method has proven efficient in Bayesian estimation of gating parameters. Furthermore, we have proposed an effective scheme for sampling the indexing parameter under various scenarios.

Our simulation results demonstrate that the Bayesian approach outperforms traditional likelihood-based inference, particularly in cases with censored data. Additionally, analysis of real data suggests that the proposed model exhibits robustness against outliers, making it a promising tool for applications involving censored and heavy-tailed data.

6 Acknowledgment

We would like to express our sincere gratitude to our supervisor, Dr. Arnab Hazra, for his invaluable guidance and support throughout this project. His expertise and constant encouragement played a key role in the success of this work. We are deeply appreciative of his contributions. We gratefully acknowledge the contributions of Elham Mirfarah, Mehrdad Naderi, Tsung-I Lin, and Wan-Lun Wang, whose work in the referenced paper served as a foundation and inspiration for this project.

7 Our Contribution

- **Finding papers:** Ahana Bose, Sneha Karmakar, Sohini Bhadra, Rishikesh Dargad
- **Data Application:** Rishikesh Dargad, Ahana Bose, Sneha Karmakar, Sohini Bhadra
- **Simulation Study:** Sohini Bhadra, Rishikesh Dargad
- **Tables and Plots:** Rishikesh Dargad
- **Report Writing:** Sneha Karmakar, Ahana Bose