LIC for Distributed Skewed Regression

Hengxin Gao, Guangbao Guo

***Abstract*—We introduce a novel distributed skewed regression that combines the flexibility of skewed distributions with the efficiency of distributed computing, effectively addressing the challenges associated with large-scale skewed datasets. Within this framework, we propose an optimal subset selection criterion named LIC. Comparative analysis with two widely used metrics demonstrates that LIC achieves superior stability and sensitivity in reducing estimation errors. In addition, we evaluate the applicability of the LIC to various skewed regression models, with experimental data further corroborating its robustness and stability.**

***Index Terms*—Skewed distribution, Distributed skewed regression, Optimal subset, LIC.**

I. Introduction

IN traditional statistical analysis, normal distribution and t-distribution are commonly employed for modeling and inference. However, many real-world datasets exhibit substantial skewness. Distributed skewed regression integrates the adaptability of skewed distributions with the computational efficiency of distributed algorithms, offering an effective solution for large-scale skewed datasets.

We explore the theoretical foundations and implementation methods of two distributed skewed regression models: distributed skew-normal regression and distributed skew-t regression. Additionally, we introduce a model selection criterion named LIC, specifically designed for distributed skewed regression.

A. Current Research Status

Since Azzalini introduced skewed distributions in 1985, they have attracted considerable attention and study. In distributed statistical learning, techniques such as averaging and partitioning are widely used for large-scale data processing. Optimal subset selection is also an effective strategy for managing large datasets. Researchers have explored the subset selection problem using the Pareto approach and developed a distributed POSS algorithm with bounded approximation guarantees.

B. Our Work

In this study, we introduce the distributed skewed regression model and propose a novel optimal subset selection criterion named LIC. We compared the MAE of the three methods: LIC, minimum information, and maximum gain matrices (LIC, opt_1, and opt_2, respectively). The results show that LIC consistently outperforms the others. We explore the stability and sensitivity of these three methods for two common skewed distributions.

Manuscript received February 18, 2025; revised July 5, 2025.

This work was supported by a grant from the National Social Science Foundation Project under project ID 23BTJ059.

Hengxin Gao is an undergraduate student from Shandong University of Technology, Zibo, China (e-mail: ghx17615749973@163.com).

Guangbao Guo is a professor from Shandong University of Technology, Zibo, China (corresponding author to provide phone:15269366362; e-mail: ggb11111111@163.com).

II. Distributed skewed regression

A. Distributed Skewed Regression

Distributed skewed regression is a method that applies skewed regression models to distributed data. It decomposes computationally intensive problems into parallelizable subproblems. The procedure consists of the following steps:

1) Model Definition: The distributed partial skew-normal regression model is defined as

$$Y_{I_k} = X_{I_k}\beta + \varepsilon_{I_k}, \varepsilon_{I_k} \sim \text{Skew} - Normal(\mu_{I_k}, \sigma, \lambda), \quad (1)$$

where $\mu_{l_k} = X_{l_k}\beta$ is the mean of the regression equation$k = 1, 2...K_n$, β is the vector of regression coefficients, $\beta = (\beta_1, \cdots, \beta_p)^T$. σ is the difference in the standard deviation and λ is the bias parameter.

The distributed partial skew-t regression model is defined as

$$Y_{I_k} = X_{I_k}\beta + \varepsilon_{I_k}, \varepsilon_{I_k} \sim \text{Skew} - t(\mu_{I_k}, \sigma, \lambda, \nu). \quad (2)$$

2) Maximum Likelihood Estimate:

a) Constructing the Likelihood Function: the likelihood function can be expressed as

$$L = \Pi_{k=1}^{K_n} f\left(Y_{I_k}\right). \quad (3)$$

The logarithmic likelihood function is obtained by taking the logarithm:

$$\ell = \sum_{k=1}^{K_n} log f(Y_{I_k}; X_{I_k}\beta, \sigma, \alpha). \quad (4)$$

b) EM Algorithm to Optimize the Parameters: The EM algorithm approximates the optimal solution step by step through the alternating execution of the E-step (expectation step) and the M-step (maximization step).

Step E (Expectation Step): In this step, the expected value of the hidden variable is calculated. For skewed distributions, the hidden variable represents a potential component of the error term.

Step M (Maximization Step): In this step, the likelihood function is maximized to update the parameters β, σ, and α.

3) Local Estimation: The local regression coefficient β_k and the skewness parameter α_k were independently calculated at each computational node using the maximum likelihood estimation (MLE) of the skewed distribution.

4) Aggregation: The local estimates from all computing nodes were averaged and aggregated to obtain a global estimate.

$$\hat{\beta} = \frac{1}{K}\Sigma_{k=1}^{K}\hat{\beta}_k, \quad (5)$$

$$\hat{\alpha} = \frac{1}{K}\Sigma_{k=1}^{K}\hat{\alpha}_k. \quad (6)$$

B. Interval Estimation for Distributed Skewed Regression

Our goal is to obtain the confidence intervals for the regression coefficients $\hat{\beta}_{I_k}$ at a given confidence level $1-\alpha$ based on Y_{I_k} at $C(Y_{I_k})$.

$$P(\hat{\beta}_{I_k} \in C(Y_{I_k})|\hat{\beta}_{I_k}) = 1-\alpha. \tag{7}$$

We define $w \in (0,1)$ as the weight function, with each function corresponding to a specific confidence domain process. This process is thoroughly described in the study by Yu and Hoff and has been applied to the Bayesian optimal function. Next, we assume that $A_w(\hat{\beta}_{I_k})$ denotes the acceptance region for each $\hat{\beta}_{I_k}$, and the function satisfies the following conditions.

$$A_w(\hat{\beta}_{I_k}) = \left\{ \overline{Y}_{I_k} : \begin{pmatrix} \hat{\beta}_{I_k} - t_{n_{I_k}-p\cdot 1-\alpha w}\hat{\sigma}_{I_k}\cdot\overline{C}_{I_k}, \hat{\beta}_{I_k} \\ -t_{n_{I_k}-p,\alpha(1-w)}\hat{\sigma}_{I_k}\cdot\overline{C}_{I_k}, \end{pmatrix} \right\} \tag{8}$$

where $\bar{C}_{I_k} = \frac{\sum_{i=1}^{n_{I_k}} C_{I_k,x_i}}{n_{I_k}}, x_i \in X_{I_k}$,because $C_{I_k,x_i} = x_i^\top (X_{I_k}^\top X_{i_k})^{-1} x_i$ is the i-th diagonal element of the matrix $x_{I_k}(X_{I_k}^\top X_{i_k})^{-1}x_{I_k}^\top$, it can be expressed as

$$diag\{X_{I_k}(X_{I_k}^\top X_{I_k})^{-1}X_{I_k}^\top\} = \left(C_{I_k,x_1}, C_{I_k,x_2}, \cdots, C_{I_k,x_{n_{I_k}}}\right)$$

The confidence interval is derived by inverting the acceptance region associated with the level of α. The confidence interval for the regression coefficient $\hat{\beta}_{I_k}$ when $w = \frac{1}{2}$ is defined as

$$C(Y_{I_k}) = \left\{ \begin{array}{l} \hat{\beta}_{Ik} : \overline{Y}_{Ik} + t_{n_{I_k}-p,\frac{\alpha}{2}}\hat{\sigma}_{Ik}\cdot\overline{C}_{Ik} \le \hat{\beta}_{Ik} \\ \le \overline{Y}_{I_k} + t_{n_{I_k}-p,1-\frac{\alpha}{2}}\hat{\sigma}_{Ik}\cdot\overline{C}_{Ik} \end{array} \right\}. \tag{9}$$

Among them: $E(\hat{\sigma}_{I_k}^2) = \sigma_{I_k}^2$ .Notes:

$$\hat{\sigma}_{I_k}^2 = \frac{1}{n_{I_k}-p}\hat{\varepsilon}_{I_k}^\top\hat{\varepsilon}_{I_k} = \frac{1}{n_{I_k}-p}Y_{I_k}^\top\left(I_{n_{I_k}\times n_{I_k}} - H_{I_k}\right)Y_{I_k}, \tag{10}$$

where $\hat{\varepsilon}_{I_k} = Y_{I_k} - \hat{Y}_{IK} = (I_{n_{I_k}\times n_{I_k}} - H_{I_k})Y_{I_k}$.

For the full-rank submatrix $X_{I_k}^\top X_{I_k}$, we have

$$H_{I_k} = X_{I_k}(X_{I_k}^\top X_{I_k})^{-1}X_{I_k}^\top. \tag{11}$$

In the special case where the matrix is not invertible, we have

$$H_{I_k} = X_{I_k}(X_{I_k}^\top X_{I_k} + \lambda I_{n\times n})^{-1}X_{I_k}^\top, \tag{12}$$

where λ is the interference term, and $I_{n\times n}$ is the original matrix of $n\times n$. Then, the shortest interval length for β_{I_k} can be obtained as follows:

$$L(C(Y_{I_k})) = 2\hat{\sigma}_{I_k}\cdot\overline{C}_{I_k}\cdot t_{n_{I_k}-p,1-\frac{\alpha}{2}}. \tag{13}$$

In addition, when n_{I_k} is large enough, we use the Z distribution instead of the t distribution to compute the length of the interval, which is also mentioned in Javanmard and Yuchen Zhang et al. The length satisfies.

$$2\hat{\sigma}_{I_k}\cdot\overline{C}_{I_k}\cdot t_{n_{I_k}-p,1-\frac{\alpha}{2}}.$$

III. LIC FOR DISTRIBUTED SKEWED REGRESSION

In the first step, for the subset sequence $\{I_k\}_{k=1}^{K_n}$, the optimal indicator subset is selected based on the shortest interval length of I_{opt}^1 , such that

$$I_{\text{opt}}^1 = \arg\min_{I_k}\left\{\hat{\sigma}_{I_k}\cdot\overline{C}_{I_k}\cdot t_{n_{I_k-1},1-\frac{\alpha}{2}}\right\}, \tag{14}$$

where $\hat{\sigma}_{I_k}$, $\overline{C}_{I_k}$ and $t_{n_{I_k-1,1-\frac{\alpha}{2}}}$ are derived from the formula $L(C(Y_{I_k})) = 2\hat{\sigma}_{I_k}\cdot\overline{C}_{I_k}\cdot t_{n_{I_k}-p,1-\frac{\alpha}{2}}$.

In the second step, it is possible to prove the least-squares estimate of β_{I_k} and the variance of $\hat{\beta}_{I_k}$.

$$\hat{\beta}_{I_k} = (X_{I_k}^\top X_{I_k})^{-1}X_{I_k}^\top Y_{I_k}, var(\hat{\beta}_{I_k}) = \hat{\sigma}_{I_k}^2(X_{I_k}^\top X_{I_k})^{-1}, \tag{15}$$

where $E(\hat{\sigma}_{I_k}^2) = \sigma_{I_k}^2$. Based on the maximization information matrix $X_{I_k}^\top X_{I_k}$, the optimal indication subset I_{opt}^2 is obtained:

$$I_{\text{opt}}^2 = \arg\max_{I_k}|X_{I_k}^\top X_{I_k}|. \tag{16}$$

The second step of the algorithm mirrors the method used in the IBOSS algorithm under the D-optimality criterion proposed by Haiying Wang. Specifically, it involves selecting a subset of i data points from the dataset containing K_n two-dimensional variables (Y_{I_k}, X_{I_k}) to maximize the following equation.

$$\delta_{\text{opt}}^D = \arg\max_{\delta}\left|\sum_{k=1}^{K_n}\delta_k X_{I_k}X_{I_k}^\top\right|, \sum_{k=1}^{K_n}\delta_k = 1, \tag{17}$$

where, δ_k is an indicator variable. When $\delta_k = 1$, the pair (Y_{I_k}, X_{I_k}) is included in the subset, whereas when $\delta_k = 0$, (Y_{I_k}, X_{I_k}) is excluded from the subset.

In the third step, to further eliminate redundant information and reduce the size of the subset, the following calculation is performed to obtain the final optimal subset.

$$I_{\text{opt}} = I_{\text{opt}}^1 \cap I_{\text{opt}}^2. \tag{18}$$

Therefore, the optimal subset $Q_{I_{\text{opt}}} = (Y_{I_{\text{opt}}}, X_{I_{\text{opt}}})$ is selected from all possible subsets $\{Q = (Y_{I_k}, X_{I_k})\}_{k=1}^{K_n}$. This criterion, which is related to the length of the interval and the information matrix, is referred to as the LIC. For this optimal subset, the shortest interval length of $\beta_{I_{\text{opt}}}$ was achieved.

$$L(C(Y_{I_{\text{opt}}})) = \hat{\sigma}_{I_{\text{opt}}}\cdot\overline{C}_{I_{\text{opt}}}\cdot t_{n_{I_{\text{opt}}}-1,1-\alpha/2}. \tag{19}$$

In the problem discussed above, we use the LIC to select the optimal subset of indications.

IV. NUMERICAL ANALYSIS

To evaluate the performance of the proposed LIC, simulations were carried out. We also analyzed the performance of two other metrics. opt_1 and opt_2 under the same conditions.

A. Prepare

By using three indicator subsets I_{opt},I_{opt}^1 and I_{opt}^2, we have $\hat{\beta}_{I_{\text{opt}}}$, $\hat{\beta}_{I_{\text{opt}}^1}$, $\hat{\beta}_{I_{\text{opt}}^2}$. The MAE was chosen as an indicator of the estimation.

B. Stability analysis

1) Analog Background: It is assumed that the error term obeys a skew-normal distribution and a skew-t distribution, we analyze the stability and sensitivity of the LIC.

The generated dataset (X, Y) varies depending on the distribution of the error term.

Case 1 (Skew-Normal distribution):

$Y_1 = X_1\beta + \varepsilon_1, \varepsilon_1 \sim Skew - Normal(5, 2, 0.05)$,

Case 2 (Skew-t distribution):

$Y_1 = X_1\beta + \varepsilon_1, \varepsilon_1 \sim Skew - t(5, 2, 0.05)$,

where X consists of (X_1, X_2) and Y consists of (Y_1, Y_2). The definitions are as

$$X_1 = (X_{ij}) \in \mathbb{R}^{n_1 \times p}, \quad X_{1j} \sim N(0, 4),$$
$$X_2 = (X_{ij}) \in \mathbb{R}^{n_2 \times p}, \quad X_{2j} \sim \text{Beta}(2, 1),$$
$$Y_1 = X_1\beta + \varepsilon_1, \quad n_1 = n - n_r,$$
$$Y_2 = X_2\beta + \varepsilon_2, \quad n_2 = n_r.$$

Note that $\beta \sim Unif(1, 5)$ and $\varepsilon = (\varepsilon_1, \varepsilon_2)$, where $\beta \sim Laplace(0, 8)$ runs our simulation.

2) Simulation Analysis: The stability of LIC under skew-normal distribution and skew-t distribution is investigated by varying the values of n and p.

When n varies over the set {2000,3000,4000,5000,6000}, {p, K, α, σ_1, σ_2, n_r} = {8, 10, 0.05, 1, 8, 10}.

When p varies over the set {8,9,10,11,12}, {n, K, α, σ_1, σ_2, n_r} = {2000, 10, 0.05, 1, 8, 10}.

Case 1: Stability analysis of LIC under the skew-normal distribution.

i. Effect of n-value on the stability of the LIC

Fig. 1 shows that under the skew-normal distribution, the MAE of all three criteria decreases as n increases, The LIC fluctuates the smallest fluctuation and achieves the best performance when $n = 3000$ and $p = 8$.

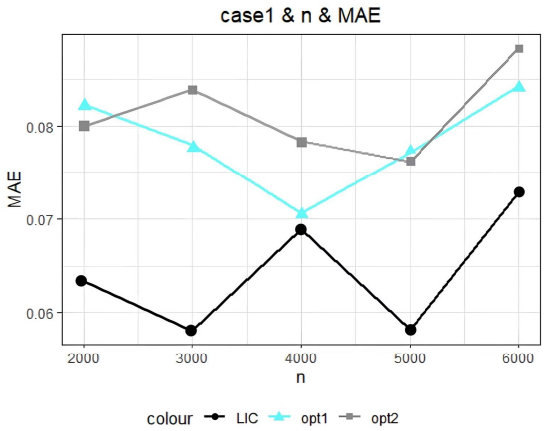

Fig. 1. Stability analysis of LIC for n-value variations under skew-normal distribution.

ii. Effect of p-value on the stability of the LIC

Fig. 2 shows that under the skew-normal distribution, the MAE of all criteria initially decreases, then increases, and finally falls again as p varies from 8 to 12. The LIC achieves the lowest MAE at $p = 12$, making it the best option.

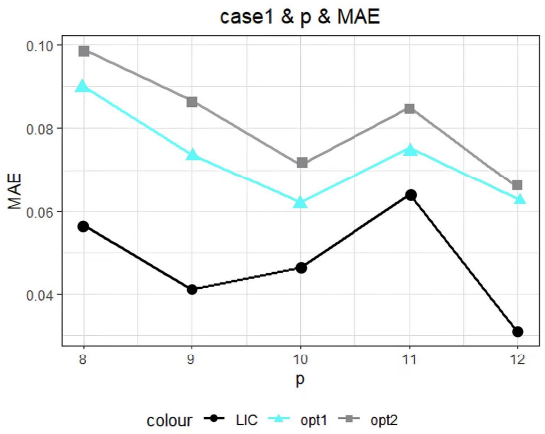

Fig. 2. Stability analysis of LIC for p-value variations under skew-normal distribution.

Case 2: This case studies the stability of the LIC under a skew-t distribution.

i. Effect of n-value on the stability of the LIC

Fig. 3 shows that under the skew-t distribution, the MAE curves of the LIC differ from opt_1 and opt_2 as n increases. Although the LIC's MAE values trend upward. it achieves its best performance with the lowest error at $n = 3000$ and $p = 8$.

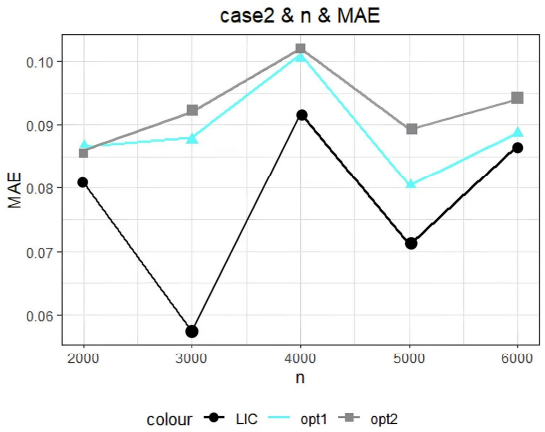

Fig. 3. Stability analysis of LIC for n-value variations under skew-t distribution.

ii. Effect of p-value on the stability of the LIC

Fig. 4 shows that under the skew-t distribution, notable performance differences are observed across criteria. The LIC achieves the lowest MAE at p=11, demonstrating its optimal performance.

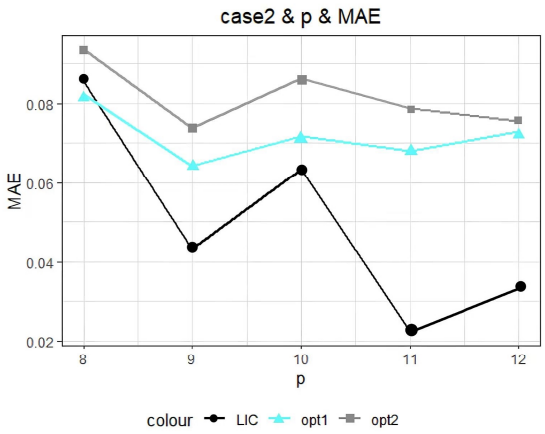

Fig. 4. Stability analysis of LIC for p-value variations under Skew-t distribution.

3) Sensitivity Analysis: The sensitivity of the LIC under skewed-normal and skewed-t distributions is examined by varying the values of K and n_r.

When K varies over the set {5,10,15,20,25}, {p, n, α, σ_1, σ_2, n_r} = {8, 2000, 0.05, 1, 8, 10}.

When n_r varies over the set {50,60,70,80,90}, {p, n, α, σ_1, σ_2, K} = {8, 2000, 0.05, 1, 8, 10}.

Case 3: This case studies the sensitivity of the LIC under skew-normal distribution conditions.

i. Effect of K-value on the sensitivity of the LIC

Fig. 5 shows that under the skew-normal distribution, the LIC maintains stable performance with remains relatively stable when K ranges from 5 to 15. However, as K increases beyond 15, variability rises, and the LIC achieves the lowest MAE at $K = 20$, indicating sensitivity to larger K values.

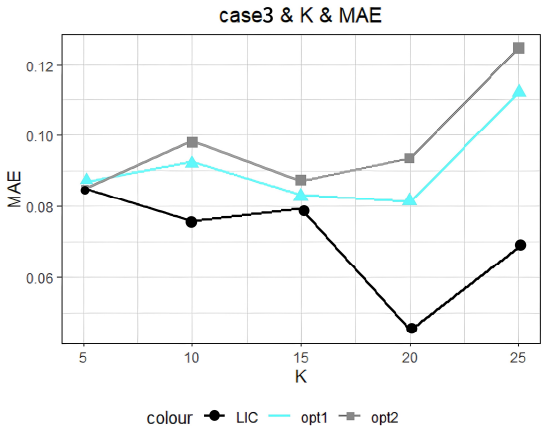

Fig. 5. Sensitivity analysis of LIC for K-value variations under skew-normal distribution.

ii. Effect of n_r-value on the sensitivity of the LIC

Fig. 6 shows that under the skew-normal distribution, the LIC is sensitive to changes in n_r, exhibiting greater fluctuation in MAE compared to opt_1 and opt_2. Despite this variability, the overall error remains relatively low.

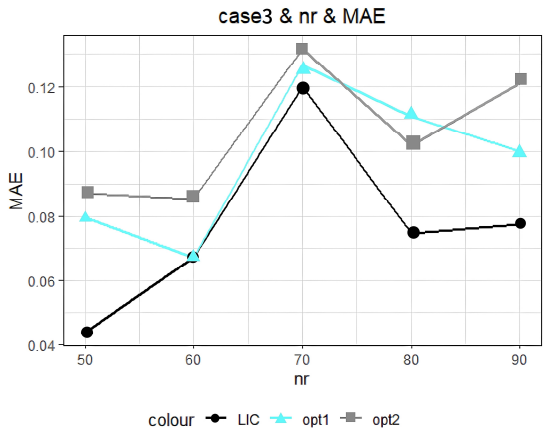

Fig. 6. Sensitivity analysis of LIC for n_r -value variations under skew-normal distribution.

Case 4: This case studies the sensitivity of the LIC under a skew-t distribution.

i. Effect of K-value on the sensitivity of the LIC

Fig. 7 shows that under the skew-t distribution, the LIC exhibits great sensitivity to the choice of K, especially in the range of 10 and 25. Nevertheless, the overall MAE remains within a low range.

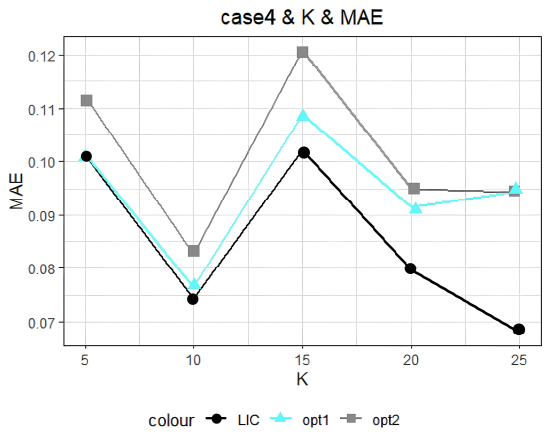

Fig. 7. Sensitivity analysis of LIC for K-value changes under skew-t distribution.

ii. Effect of n_r-value on the sensitivity of the LIC

Fig. 8 shows that under the skew-t distribution, the LIC is sensitive to changes in n_r. The MAE of the LIC fluctuates with changes in n_r, with notable local peaks around $n_r = 60$. This indicates that while the LIC's performance can be affected by n_r changes, it generally maintains a low error level.

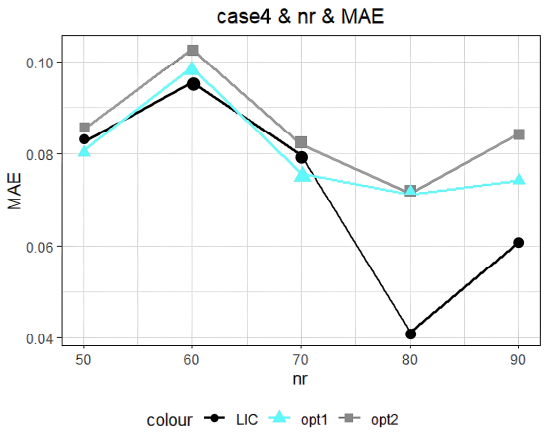

Fig. 8. Sensitivity analysis of LIC for n_r-value variations under skew-t distribution.

V. Conclusion

The proposed distributed skewed regression and the LIC offer an efficient and reliable solution for large-scale skewed data. Its applicability to various distributions has been confirmed. Future research will focus on practical applications and its potential in complex data scenarios.

References

[1] Azzalini, A. A class of distributions which includes the normal ones. *Scand. J. Statist.*, 12, 171–178.

[2] Azzalini, A. with the collaboration of Capitanio, A. *The Skew-Normal and Related Families*. Cambridge University Press, IMS Monographs series.

[3] Azzalini, A. and Capitanio, A. Distributions generated by perturbation of symmetry with emphasis on a multivariate skew-t distribution. *J. Roy. Statist. Soc. B*, 65, 367–389.

[4] Jamalizadeh, A., Khosravi, M., and Balakrishnan, N. Recurrence relations for distributions of a skew-t and a linear combination of order statistics from a bivariate-t. *Comp. Statist. Data An.*, 53, 847–852.

[5] Fieller, N. J., Flenley, E. C., and Olbricht, W. Statistics of Particle Size Data. *Applied Statistics*, 41, 127–146.

[6] G. Guo, Q. Wang, J. Allison, and G. Qian. Accelerated distributed expectation-maximization algorithms for the parameter estimation in multivariate Gaussian mixture models. *Applied Mathematical Modelling*, 137, 115709.

[7] G. Guo, H. Song, and L. Zhu. The COR criterion for optimal subset selection in distributed estimation. *Statistics and Computing*, 34, 163–176.

[8] G. Guo, M. Yu, and G. Qian. ORKM: Online regularized K-means clustering for online multi-view data. *Information Sciences*, 680, 121133.

[9] Q. Wang, G. Guo, G. Qian, and X. Jiang. Distributed online expectation-maximization algorithm for Poisson mixture model. *Applied Mathematical Modelling*, 124, 734–748.

[10] G. Guo, C. Wei, and G. Qian. Sparse online principal component for the parameter estimation in factor model. *Computational Statistics*, 38(2), 1095–1116.

[11] G. Guo, G. Qian, and L. Zhu. A scalable quasi-newton estimation algorithm for dynamic generalized linear models. *Journal of Nonparametric Statistics*, 34, 917–939.

[12] G. Guo, Y. Sun, G. Qian, and Q. Wang. LIC criterion for optimal subset selection in distributed interval estimation. *Journal of Applied Statistics*, 50(9), 1900–1920.

[13] G. Guo and W. Zhao. Schwarz method for financial engineering. *Journal of Computational Mathematics*, 39(4), 538–555.

[14] G. Guo, J. Allison, and L. Zhu. Bootstrap maximum likelihood for quasi-stationary distributions. *Journal of Nonparametric Statistics*, 31(1), 64–87.

[15] W. You, Z. Yang, G. Guo, X.-F. Wan, and G. Ji. Prediction of DNA-binding proteins by interaction fusion feature representation and selective ensemble. *Knowledge-Based Systems*, 163, 598–610.

[16] G. Guo and W. Zhao. Schwarz methods for quasi stationary distributions of Markov chains. *Calcolo*, 49, 21–39.

[17] G. Guo and S. Lin. Schwarz method for penalized quasi-likelihood in generalized additive models. *Communications in Statistics-Theory and Methods*, 39, 1847–1854.

[18] G. Guo, W. You, G. Qian, et al. Parallel maximum likelihood estimator for multiple linear regression models. *Journal of Computational and Applied Mathematics*, 273, 251–263.

[19] G. Guo, W. Shao, L. Lin, et al. Parallel tempering for dynamic generalized linear models. *Communications in Statistics-Theory and Methods*, 45(21), 6299–6310.

[20] G. Guo. Parallel statistical computing for statistical inference. *Journal of Statistical Theory and Practice*, 6(3), 536–565.

[21] G. Guo, G. Qian, L. Lin, et al. Parallel inference for big data with the group Bayesian method. *Metrika*, 84, 225–243.

[22] G. Guo and L. Lin. Parallel bootstrap and optimal subsample lengths in smooth function models. *Communications in Statistics-Simulation and Computation*, 45(6), 2208–2231.

[23] G. Guo, Y. Sun, and X. Jiang. A partitioned quasi-likelihood for distributed statistical inference. *Computational Statistics*, 35, 1577–1596.

[24] W. Shao and G. Guo. Multiple-Try simulated annealing algorithm for global optimization. *Mathematical Problems in Engineering*, 2018(1), 9248318.

[25] W. Shao, G. Guo, G. Zhao, et al. Simulated annealing for the bounds of Kendall's τ and Spearman's ρ. *Journal of Statistical Computation and Simulation*, 84(12), 2688–2699.

[26] W. Shao, G. Guo, F. Meng, et al. An efficient proposal distribution for Metropolis-Hastings using a B-splines technique. *Computational Statistics & Data Analysis*, 57(1), 465–478.

[27] G. Guo. A block bootstrap for quasi-likelihood in sparse functional data. *Statistics*, 54(5), 909–925.

[28] L. Song and G. Guo, "Full information multiple imputation for linear regression model with missing response variable," IAENG International Journal of Applied Mathematics, vol.54, no.1, pp77-81, 2024.

[29] G. Guo and S. Lin. Schwarz method for penalized quasi-likelihood in generalized additive models. *Communications in Statistics-Theory and Methods*, 39(10), 1847–1854.

[30] G. Guo. Finite difference methods for the BSDEs in finance. *International Journal of Financial Studies*, 6(1), 26.

[31] G. Guo and G. Qian. Optimal subset selection for distributed local principal component analysis. *Physica A: Statistical Mechanics and its Applications*, 658, 130308.

[32] D. Chang and G. Guo. LIC: An R package for optimal subset selection for distributed data. *SoftwareX*, 28, 101909.

[33] G. Guo, R. Niu, G. Qian, et al. Trimmed scores regression for k-means clustering data with high-missing ratio. *Communications in Statistics-Simulation and Computation*, 53(6), 2805–2821.

[34] G. Guo. Taylor quasi-likelihood for limited generalized linear models. *Journal of Applied Statistics*, 48(4), 669–692.

[35] G. Guo, W. You, L. Lin, et al. Covariance matrix and transfer function of dynamic generalized linear models. *Journal of Computational and Applied Mathematics*, 296, 613–624.

[36] G. Jing and G. Guo. TLIC: An R package for the LIC for T distribution regression analysis. *SoftwareX*, 30, 102132.

[37] S. Liu and G. Guo. LFM: An R package for Laplace factor model. *SoftwareX*, 30, 102133.

[38] Y. Li and G. Guo, "Distributed monotonic overrelaxed method for random effects model with missing response," IAENG International Journal of Applied Mathematics, vol.54, no.2, pp205-211, 2024.

[39] Y. Li and G. Guo, "General unilateral loading estimation," Engineering Letters, vol.32, no.1, pp72-76, 2024.

[40] C. Zhang and G. Guo, "The optimal subset estimation of distributed redundant data," IAENG International Journal of Applied Mathematics, vol.55, no.2, pp270-277, 2025.

[41] Q. Liu and G. Guo, "Distributed estimation of redundant data," IAENG International Journal of Applied Mathematics, vol.55, no.2, pp332-337, 2025.

[42] J. Li and G. Guo, "An optimal subset selection algorithm for distributed hypothesis test," IAENG International Journal of Applied Mathematics, vol.54, no.12, pp2811-2815, 2024.

[43] G. Jing and G. Guo, "Student LIC for distributed estimation," IAENG International Journal of Applied Mathematics, vol.55, no.3, pp575-581, 2025.

[44] Guangbao Guo, Yue Sun, Guoqi Qian, and Qian Wang. LIC: The LIC Criterion for Optimal Subset Selection. URL: https://CRAN.R-project.org/package=LIC.

[45] Guangbao Guo, Haoyue Song, and Lixing Zhu. COR: The COR for Optimal Subset Selection in Distributed Estimation. URL: https://CRAN.R-project.org/package=COR.

[46] Guangbao Guo, Guoqi Qian, Yixiao Liu, and Haoyue Song. DLPCA: The Distributed Local PCA Algorithm. URL: https://CRAN.R-project.org/package=DLPCA.
[47] Guangbao Guo, Haoyue Song, and Lixing Zhu. ISR: The Iterated Score Regression-Based Estimation Algorithm. URL: https://CRAN.R-project.org/package=ISR.
[48] Qian Wang, Guangbao Guo, and Guoqi Qian. DEM: The Distributed EM Algorithms in Multivariate Gaussian Mixture Models. URL: https://CRAN.R-project.org/package=DEM.
[49] Qian Wang, Guangbao Guo, and Guoqi Qian. DOEM: The Distributed Online Expectation Maximization Algorithms to Solve Parameters of Poisson Mixture Models. URL: https://CRAN.R-project.org/package=DOEM.
[50] Guangbao Guo, Chunjie Wei, and Guoqi Qian. SOPC: The Sparse Online Principal Component Estimation Algorithm. URL: https://CRAN.R-project.org/package=SOPC.
[51] Guangbao Guo, Miao Yu, Haoyue Song, and Ruiling Niu. ORKM: The Online Regularized K-Means Clustering Algorithm. URL: https://CRAN.R-project.org/package=ORKM.
[52] Chunjie Wei and Guangbao Guo. OPC: The Online Principal Component Estimation Method. URL: https://CRAN.R-project.org/package=OPC.
[53] Guangbao Guo and Yaping Li. DLEGFM: Distributed Loading Estimation for General Factor Model. URL: https://CRAN.R-project.org/package=DLEGFM.
[54] Guangbao Guo and Yu Li. DIRMR: Distributed Imputation for Random Effects Models with Missing Responses. URL: https://CRAN.R-project.org/package=DIRMR.
[55] Guangbao Guo and Liming Song. DLMRMV: Distributed Linear Models with Response Missing Variables. URL: https://CRAN.R-project.org/package=DLMRMV.
[56] Guangbao Guo and Jiarui Li. pql: A Partitioned Quasi-likelihood for Distributed Statistical Inference. URL: https://CRAN.R-project.org/package=pql.
[57] Guangbao Guo and Jiarui Li. FPCdpca: The FPCdpca Criterion on Distributed Principal Component Analysis. URL: https://CRAN.R-project.org/package=FPC.
[58] Guangbao Guo and Jiarui Li. PPCDT: An Optimal Subset Selection for Distributed Hypothesis Testing. URL: https://CRAN.R-project.org/package=PPCDT.
[59] Guangbao Guo and Ruiling Niu. DTSR: Distributed Trimmed Scores Regression for Handling Missing Data. URL: https://CRAN.R-project.org/package=DTSR.
[60] Guangbao Guo and Yaxuan Wang. LLIC: Likelihood Criterion (LIC) Analysis for Laplace Regression Model. URL: https://CRAN.R-project.org/package=LLIC.
[61] Guangbao Guo and Di Chang. Dogoftest: Distributed Online Goodness-of-Fit Tests for Distributed Datasets. URL: https://CRAN.R-project.org/package=Dogoftest.
[62] Guangbao Guo and Yu Jin. SFM: A Package for Analyzing Skew Factor Models. URL: https://CRAN.R-project.org/package=SFM.
[63] Beibei Wu and Guangbao Guo. TFM: Sparse Online Principal Component for TFM. URL: https://CRAN.R-project.org/package=TFM.
[64] Guangbao Guo and Guofu Jing. TLIC: The LIC for T Distribution Regression Analysis. URL: https://CRAN.R-project.org/package=TLIC.
[65] Guangbao Guo and Siqi Liu. LFM: Laplace Factor Model Analysis and Evaluation. URL: https://CRAN.R-project.org/package=LFM.