Research on Distributed Redundant Data Estimation Based on LIC

Di Chang, Guangbao Guo

***Abstract*—The problem of data redundancy in distributed storage has become increasingly pronounced, posing significant challenges for the estimation of target variables. This study introduces a distributed redundant data estimation method that employs the LIC criterion. Through simulation, the method's predictive accuracy is rigorously estimated, and its stability and sensitivity are thoroughly investigated. Results demonstrate the method's effectiveness in extracting valuable information from redundant distributed data. By identifying the optimal data subset, it enhances data quality and boosts efficiency, making it a potent strategy for tackling data analysis challenges inherent in big data environments.**

***Index Terms*—data redundancy, LIC criterion, stability and sensitivity, optimal data subset.**

I. Introduction

WITH the continuous advancement of science and technology, data collection and storage capabilities have greatly improved. This has resulted in the widespread use of distributed systems as the main method for processing and storing large amounts of data. This method involves splitting data into subsets. It processes the subsets at the same time across many computing nodes. This greatly boosts data processing efficiency. Redundant information is common in big datasets. It causes problems like wasteful storage use, slower data processing, and higher costs. These problems are major barriers to good data analysis and processing.

Traditional methods for redundant data estimation usually involve centralized data processing. However, they may hit efficiency bottlenecks when dealing with large-scale data. To address this challenge, the paper proposes a method for redundant data estimation. It estimates optimal subsets using the LIC criterion. The LIC criterion is an efficient data compression technique. It can reduce data size by removing redundant information. Leveraging the LIC criterion enables us to compress data and select the optimal subset.

A. Current Research Status

Currently, significant progress has been made in the research on distributed redundant data. In the realm of distributed statistical inference, Guo et al. [1] introduced a parallel method, which has enhanced the efficiency of statistical inference. Guo et al. [2] devised an optimization program for distributed interval estimation problems, leading to improved estimation accuracy. See also Wang et al. [3], Li et al. [4], Song et al. [5], Guo et al. [6], Huang et al. [7], Wang et al. [8], Qian et al. [9], Guo et al. [10], Battey et al. [11], Minsker et al. [12], Mirzasoleiman et al. [13], Zhang et al. [14], Cheng et al. [15] and Guo et al. [16]–[22]. These methodologies collectively provide effective statistical inference tools for handling large-scale distributed data.

Manuscript received January 17, 2024; revised December 8, 2024.

This work was supported by the National Social Science Foundation of China under Grant 23BTJ059, the Natural Science Foundation of Shandong under Grant ZR2020MA022, and the National Statistical Research Program under Grant 2022LY016.

Di Chang is a postgraduate student at the School of Mathematics and Statistics, Shandong University of Technology, Zibo, 255000, China (e-mail: Cd08210122@163.com).

Guangbao Guo is a professor at the School of Mathematics and Statistics, Shandong University of Technology, Zibo, 255000, China (corresponding author to provide phone:15269366362; e-mail: ggb11111111@163.com).

B. Our Work

In this paper, we introduce the innovative LIC criterion, designed to manage redundant data and identify the optimal subset. We evaluate the performance of three subset selection methods—LIC criterion, Minimum Information Matrix, and Maximum Gain Matrix —using MAE and MSE as comparative metrics. This analysis aims to validate the LIC criterion as the superior method for subset selection. The paper elaborates on the foundational theory underpinning the LIC criterion and presents simulation experiments to assess the estimation methods' accuracy and robustness. Furthermore, we investigate the stability and sensitivity of the three methods across three prevalent redundant data distribution functions, thereby substantiating the efficacy of the LIC criterion. The primary advantage of this algorithm is its enhancement of existing estimation techniques; it achieves this by reducing subset length without compromising accuracy, concurrently enhancing work efficiency.

II. Theorem

Condition 1. (KKT conditions) Assuming that $\widehat{\beta}^* \in \mathbb{R}^p$ signifies the genuine value of the estimator $\widehat{\beta}$, and $\widehat{k} = \{b \in \mathbb{R}^p$: $\|b\|_\infty \leq 1 \ and \ b_j = sign[\widehat{\beta}_j], \ if \ \widehat{\beta}_j \neq 0\}$ denotes a sparse boundary condition, we have $-2X_{I_{opt}^T}(Y_{I_{opt}} - X_{I_{opt}}\widehat{\beta}^*) + r\widehat{k} = 0_p$.

Condition 2. The Gram matrix $X_{I_{opt}}^T X_{I_{opt}} \in \mathbb{R}^{p\times p}$ is revertible, the ordered eigenvalues are $m_p \geq \cdots \geq m_1 > 0$, so that ordered eigenvalues of the inverse matrix $(X_{I_{opt}}^T X_{I_{opt}})^{-1}$are $1/m_1 \geq \cdots \geq 1/m_p > 0$, which satisfied $\|(X_{I_{opt}}^T X_{I_{opt}})^{-1}\alpha\| \leq \|\alpha\|_2/m_1$.

Theorem 1. Suppose that the sample data set is sparse. Let the Lasso estimator is $\widehat{\beta}_{lassso}$, the truth estimator is $\widehat{\beta}^*$ and the optimal estimator is $\widehat{\beta}_{Iopt}$. Under Condition 1-2, we have $\|X_{lopt}\widehat{\beta}^* - X_{lopt}\widehat{\beta}_{lopt}\|_2^2 \leq \frac{\lambda^2 p}{4m_1}$.

Proof. Based on data sub-matrix Q_{I_k}, suppose Lasso optimal regression model of the form $Y_{I_{opt}} = X_{Iopt}\beta_{I_{opt}} + u_{I_{opt}}$, $u_{I_{opt}} \sim N(0, \sigma^2 I_{n_{I_{opt}}\times n_{I_{opt}}})$. Denote I_{opt} is indicator set function, $X_{I_{opt}}$ is an $n_{I_{opt}}\times\ p$ optimal submatrix of X with $n_{I_{opt}} \geq p$, σ^2 is the sample's unknown variance and $u_{\mathrm{I_{opt}}}$ is an error sub-vecton with $E\left(u_{I_{opt}}\right) = 0$ and $\mathrm{Var}\left(u_{I_{opt}}\right) = \sigma^2 I_{n_{I_{opt}}\times n_{I_{opt}}}$. That the optimal Gram

matrix $X_{I_{opt}}^T X_{I_{opt}}$ and the matrix eigenvalues are denoted $m_i (i = 1, \ldots, p)$. Let estimator $\beta = (\beta_1, \ldots, \beta_p)^T$ be a p-dimensional vector, certain tuning parameters $\lambda \in [0, \infty)$. The lasso's KKT condition

$$-2X_{I_{opt}}^T \left(y - X_{I_{opt}} \widehat{\beta}_{lasso} \right) + \lambda \widehat{k} = 0_p.$$

When the data is sparse, the Lasso estimator approximates the population truth estimator of the data ($\widehat{\beta}_{lasso} \approx \widehat{\beta}^*$). Defined the optimal estimator is

$$\widehat{\beta}_{Iopt} = \left(X_{I_{opt}}^T X_{I_{opt}} \right)^{-1} X_{I_{opt}}^T Y_{I_{opt}}.$$

Adjust by equation, such that

$$\widehat{\beta}^* = \left(X_{I_{opt}}^T X_{I_{opt}} \right)^{-1} X_{I_{opt}}^T Y_{I_{opt}} - \frac{\lambda}{2} \left(X_{I_{opt}}^T X_{I_{opt}} \right)^{-1} \widehat{k}.$$

We then find

$$\begin{aligned}
&\|X_{I_{\text{opt}}} \widehat{\beta}^* - X_{I_{\text{opt}}} \widehat{\beta}_{I_{\text{opt}}}\|_2^2 \\
&= \| - \frac{\lambda}{2} X_{I_{\text{opt}}} (X_{I_{\text{opt}}}^T X_{I_{\text{opt}}})^{-1} \widehat{k}\|_2^2 \\
&= \frac{\lambda^2 \|X_{I_{\text{opt}}} (X_{I_{\text{opt}}}^T X_{I_{\text{opt}}})^{-1} \widehat{k}\|_2^2}{4} \\
&= \frac{\lambda^2 \left(X_{I_{\text{opt}}} (X_{I_{\text{opt}}}^T X_{I_{\text{opt}}})^{-1} \widehat{k} \right)^T X_{I_{\text{opt}}} (X_{I_{\text{opt}}}^T X_{I_{\text{opt}}})^{-1} \widehat{k}}{4} \\
&= \frac{\lambda^2 \widehat{k}^T \left((X_{I_{\text{opt}}}^T X_{I_{\text{opt}}})^{-1} \right)^T X_{I_{\text{opt}}}^T X_{I_{\text{opt}}} (X_{I_{\text{opt}}}^T X_{I_{\text{opt}}})^{-1} \widehat{k}}{4} \\
&= \frac{\lambda^2 \widehat{k}^T \left((X_{I_{\text{opt}}}^T X_{I_{\text{opt}}})^{-1} \right)^T \widehat{k}}{4} \\
&= \frac{\lambda^2 \widehat{k}^T (X_{Iopt}^T X_{Iopt})^{-1} \widehat{k}}{4} \\
&\leq \frac{\lambda^2 \parallel \widehat{k} \parallel_2 \parallel (X_{I_{\text{opt}}}^T X_{I_{\text{opt}}})^{-1} \widehat{k} \parallel_2}{4} \\
&\leq \frac{\lambda^2 \parallel \widehat{k} \parallel_2^2}{4m_1}.
\end{aligned}$$

Under Condition 1 and the ℓ_2-norm definition can be obtained $|\widehat{k}_j| \leq 1$ since $\widehat{k} \in \partial \|\widehat{\beta}_{lasso}\|_1$ and $\|\widehat{k}\|_2^2 = \sum_{j=1}^p |\widehat{k}_j|^2 \leq \sum_{j=1}^p 1 = p$.

From Lasso sparsity boundary conditions, we can conclude $\|X_{I_{opt}} \beta^* - X_{I_{opt}} \widehat{\beta}_{I_{opt}}\|_2^2 \leq \frac{\lambda^2 p}{4m_1}$. □

It can be seen that the boundary between the true value $\widehat{\beta}^*$ of the sparse data estimator and the optimal value $\widehat{\beta}_{I_{opt}}$ of the simulation estimator is less than or equal to the constant $\frac{\lambda^2 p}{4m_1}$.

III. Steps

For convenient reference, we delineate the selection steps as follows:

i: Generating simulated data sets for each distribution using R software.

ii: Employing an algorithm rooted in LIC distributed redundant data estimation to process this data.

iii: Employing evaluation metrics to assess the performance of the three algorithms and documenting the results for LIC, Lopt, and Iopt.

iv: Evaluating the performance of the LIC criterion and identifying the optimal subset through visualization examination.

v: Analyzing the characteristics and advantages of the optimal subset and discussing its practical applicability.

IV. Simulation Study

A. Simulation preparation

The (X, Y) is from $Y_i = X_i \beta_i + u_i$, where $u_i \sim N\left(0, \sigma_i^2 I_{n \times n}\right)$ for $i = 1, 2$. In this simulation, we construct X to comprise (X_1, X_2), and Y consists of (Y_1, Y_2). Define as

$$\begin{aligned}
Y_1 &= X_1 \beta + u_1, n_1 \in \{1, \ldots, n - n_r\}, \\
X_1 &= \left(X_{1ij}\right) \in \mathbb{R}^{n_1 \times p}, X_{1ij} \sim N(0, 2). \\
Y_2 &= X_2 \beta + u_2, n_2 \in \{1, \ldots, n_r\}, \\
X_2 &= \left(X_{2ij}\right) \in \mathbb{R}^{n_2 \times p}, X_{2ij} \sim F(X).
\end{aligned}$$

where $\beta \sim Unif(0, 3)$, $u \sim (u_1, u_2)$, $u_1 \sim N\left(0, \sigma_1^2\right)$ and $u_2 \sim N\left(0, \sigma_2^2\right)$.

In evaluating the performance of the LIC criterion, a comprehensive set of indicators is utilized. During the data simulation process, attention is given to the MSE and MAE as measures of prediction accuracy. These metrics quantify the disparities between true values and estimated values. Specifically, MSE and MAE, which relate to prediction errors, are defined as follows:

$$MSE = E(Y_0 - \widehat{Y})^2, \ MAE = E|Y_0 - \widehat{Y}|.$$

Typically, larger values of MAE and MSE indicate poorer model fitting and prediction accuracy.

The primary objective of this experiment is to generate simulated data by varying the control parameters: $\{n, p, K, n_r\}$. This is done under conditions where the redundancy distribution adheres to the uniform, geometric, and chi-square distributions. The aim is to observe and analyze the numerical variations of the LIC criterion, thereby providing a comprehensive assessment of its distinctive features and practical performance.

B. Stability

For $(\alpha, \sigma_1, \sigma_2, K, n_r) = (0.01, 3, 5, 10, 50)$, adjust the values of $\{n, p\}$ to observe changes in simulation results and determine the optimal parameters.

Case1. $X_2 = \left(X_{2ij}\right) \in \mathbb{R}^{n_2 \times p}, X_{2ij} \sim Unif(0, 3)$.

• ***Scenario I:*** Setting $n = (1000, 2000, 3000, 4000, 5000)$ with $p = 8$.

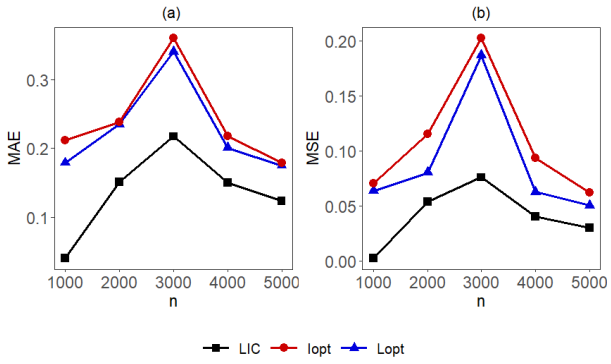

Fig. 1. The values of MAE and MSE with $n = (1000, 2000, 3000, 4000, 5000)$ and $p = 8$.

• ***Scenario II:*** Setting $p = (8, 9, 10, 11, 12)$ with $n = 3000$.

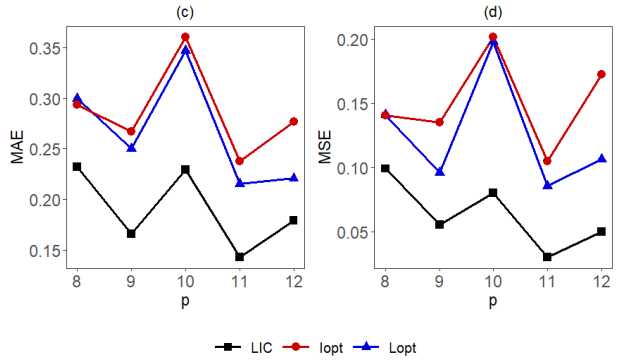

Fig. 2. The values of MAE and MSE with $p = (8, 9, 10, 11, 12)$ and $n = 3000$.

The stability of the LIC criterion under the Uniform distribution is explored. Experiments are designed with variations in the parameters n and p. A dataset of 1000 data points is generated, and the LIC criterion value for each data point is computed sequentially. The simulated results are analyzed, as shown in Fig.1 and Fig.2, with detailed considerations based on the standard values of LIC for each data point.

A comparative analysis reveals a notably similar trend in the variations of MAE and MSE. The impact of changes in n and p on the outcomes is examined, and it is observed that the LIC criterion has the lowest MAE and MSE values when compared to the Optimal Information (Iopt) and Optimal Gain Matrix selection algorithms. This observation strongly indicates the significant advantage of the LIC criterion's stability.

With a larger sample size n, the MAE and MSE curves under the LIC criterion at first rise and then fall. As n increases from 3000 to 4000, the MAE under the LIC criterion notably falls from 0.21749246 to 0.1502028, and the MSE also drops from 0.07605684 to 0.040227921.

In the variation of dimensionality p, certain patterns are observed. Notably, when the dimensionality reaches 10, the MSE peaks at 0.08007556, and the MAE also attains a relatively high value of 0.2294404. This observation offers crucial insights into the behavior of the metrics as dimensionality increases.

Case 2. $X_2 = (X_{2ij}) \in \mathbb{R}^{n_2 \times p}, X_{2ij} \sim \chi^2(20)$.

• ***Scenario I:*** Setting $n = (1000, 2000, 3000, 4000, 5000)$ with $p = 8$.

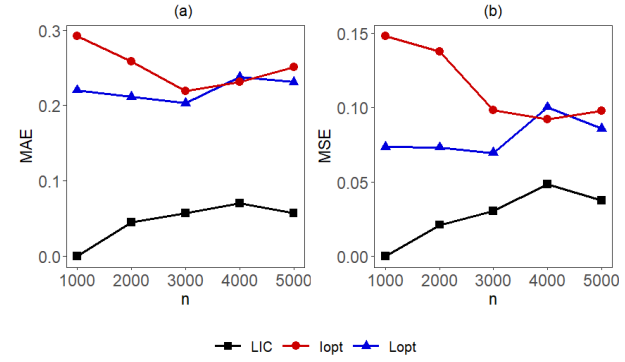

Fig. 3. The values of MAE and MSE with $n = (1000, 2000, 3000, 4000, 5000)$ and $p = 8$.

• ***Scenario II:*** Setting $p = (8, 9, 10, 11, 12)$ with $n = 3000$.

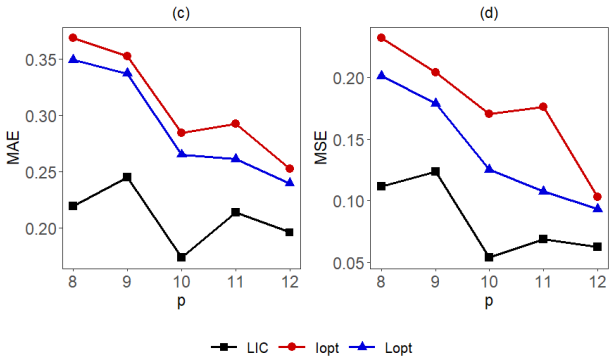

Fig. 4. The values of MAE and MSE with $p = (8, 9, 10, 11, 12)$ and $n = 3000$.

The simulated results in Fig.3 and Fig.4 are analyzed. Under the LIC criterion, MAE and MSE stayed consistently lower than the other two methods. This shows good stability. Also, the trends in variation for these two metrics are highly consistent. This further shows the LIC criterion's reliability with such data.

Specifically, as shown in Fig.3, as the sample size n grows from 1000, the values under the LIC criterion remain at extremely low levels (1.21E-05 and 2.95E-05). Then, they are rising. When n reachs 4000, both metrics reach their peak values but then exhibit a decreasing trend. This pattern of variation suggests that under the condition of $n = 1000$, the LIC criterion demonstrates optimal performance.

Similarly, as shown in Fig.4, as the dimensionality p gradually increases from 8, the MAE and MSE values under the LIC criterion generally show a decreasing trend. When $p = 10$, both MAE and MSE reached their minimum values (0.1738802 and 0.0541217, respectively), indicating that the performance of the LIC criterion achieves its optimum at this dimensionality.

Case 3. $X_2 = (X_{2ij}) \in \mathbb{R}^{n_2 \times p}$, $X_{2ij} \sim Geom(0.6)$.

• ***Scenario I:*** Setting $n = (1000, 2000, 3000, 4000, 5000)$ with $p = 8$.

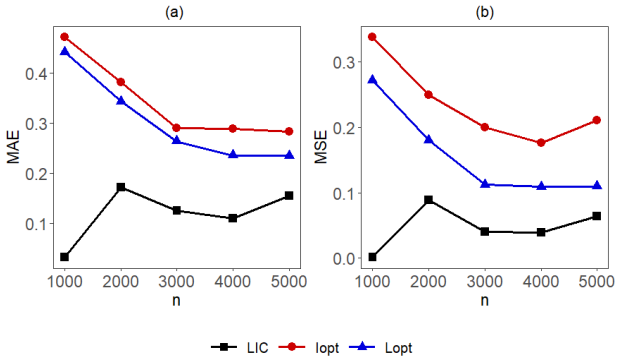

Fig. 5. The values of MAE and MSE with $n = (1000, 2000, 3000, 4000, 5000)$ and $p = 8$.

• ***Scenario II:*** Setting $p = (8, 9, 10, 11, 12)$ with $n = 4000$.

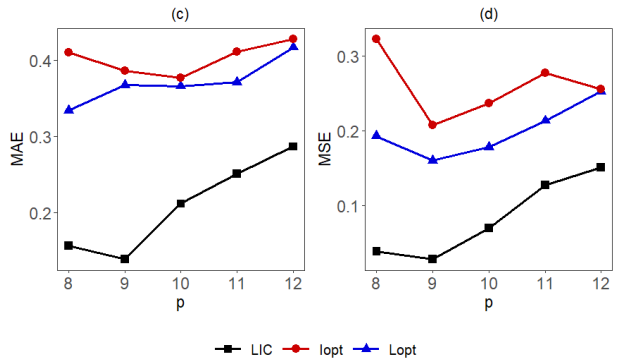

Fig. 6. The values of MAE and MSE with $p = (8, 9, 10, 11, 12)$ and $n = 4000$.

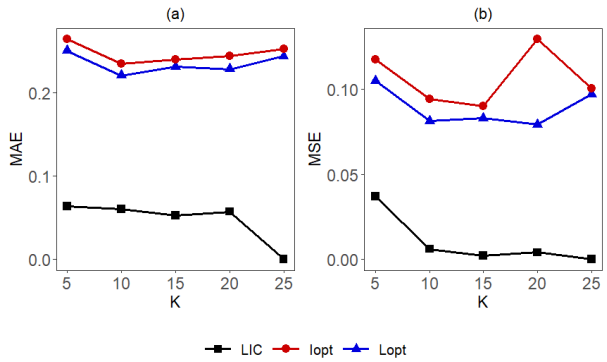

Fig. 7. The values of MAE and MSE with $K = (5, 10, 15, 20, 25)$ and $n = 6000$.

In this simulation, the stability of the geometric distribution under the LIC criterion is investigated. The focus is on selecting parameters n and p with a probability of 0.6. A comprehensive random sampling strategy is employed during the simulation process, and various combinations of n and p values are extensively tested. For each simulation, the best subset selected by the LIC criterion is recorded, along with the calculation of related performance metrics.

The trends in the MAE and MSE under the LIC criterion remained consistent, as shown in Fig.5 and Fig.6. In comparison to the Lopt and Iopt criteria, the LIC criterion results in smaller MAE and MSE values, indicating its advantages in this particular application.

Fig.5 shows that as n grew from 1000 to 2000, MAE rises from its lowest value of 0.03320847 to a peak of 0.17198431, while MSE rose from its lowest value of 0.001390464 to 0.08883672. However, as n continued to increase, the error values gradually decrease and reach their minimum at $n = 4000$, with values of 0.1109695 and 0.0386918. Similarly, Fig.6 illustrats the impact of p values on the errors. When dimensionality p is set to 9, both MAE and MSE reach their minimum values, specifically 0.1392464 and 0.02838924.

C. Sensitivity

Following initial analysis, we observe that the LIC criterion's sensitivity is substantially influenced by the critical parameters K and n_r. Additional simulations are conducted to investigate their impact on criterion performance, especially with redundant data from three key distributions. Varying K and n_r helps assess how the LIC criterion's sensitivity changes.

Case 1. $X_2 = (X_{2ij}) \in \mathbb{R}^{n_2 \times p}, X_{2ij} \sim Unif(0, 3)$ with $(\alpha, \sigma_1, \sigma_2, p, n_r) = (0.01, 3, 5, 8, 50)$.

- ***Scenario I:*** Setting $K = (5, 10, 15, 20, 25)$ with $n = 6000$.
- ***Scenario II:*** Setting $n_r = (50, 60, 70, 80, 90)$ with $n = 3000$.

The sensitivity experiment in Fig.7 is being thoroughly analyzed. In the setting, the error indicators MAE and MSE are much higher at lower block numbers K. This indicates that the LIC criterion does not improve performance as expected. However, as the K value gradually increases, the error values begin to decrease, indicating an enhancement in model performance. This trend is significant and consistent. Disregarding the influence of endpoint effects, a significant inflection point is identified at $K = 15$. At this point, both MAE and MSE reach their respective minimum values of $1.13E-05$ and $1.53E-04$. Subsequently, as the K value further increases, both error indicators show an upward trend, indicating a decline in model performance.

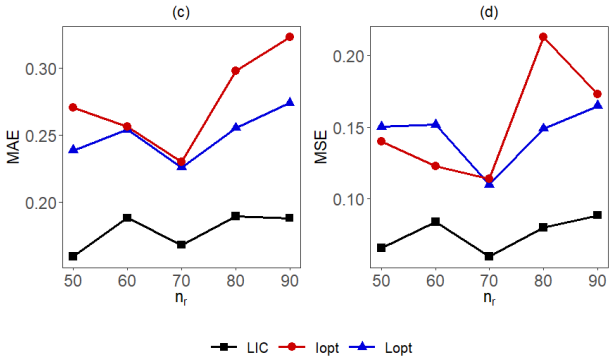

Fig. 8. The values of MAE and MSE with $n_r = (50, 60, 70, 80, 90)$ and $n = 3000$.

In Fig.8, it is seen that as the number of abnormal samples n_r grows, MAE and MSE initially increase and then decrease. Specifically, at $n_r = 70$, both reach their minimum values of 0.1683721 and 0.05974087, respectively.

Case 2. $X_2 = (X_{2ij}) \in \mathbb{R}^{n_2 \times p}, X_{2ij} \sim \chi^2(20)$ with $(\alpha, \sigma_1, \sigma_2, p, n_r) = (0.01, 3, 5, 8, 60)$.

- ***Scenario I:*** Setting $K = (5, 10, 15, 20, 25)$ with $n = 6000$.

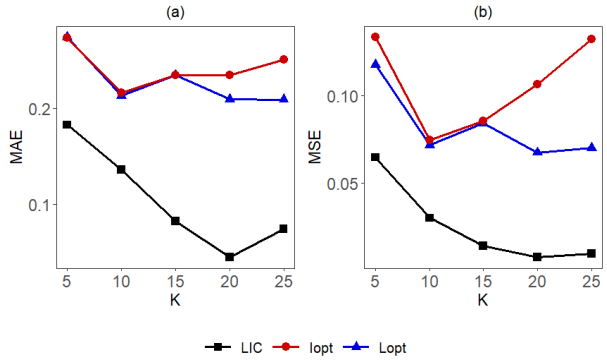

Fig. 9. The values of MAE and MSE with $K = (5, 10, 15, 20, 25)$ and $n = 6000$.

- ***Scenario II:*** Setting $n_r = (50, 60, 70, 80, 90)$ with $n =$

3000.

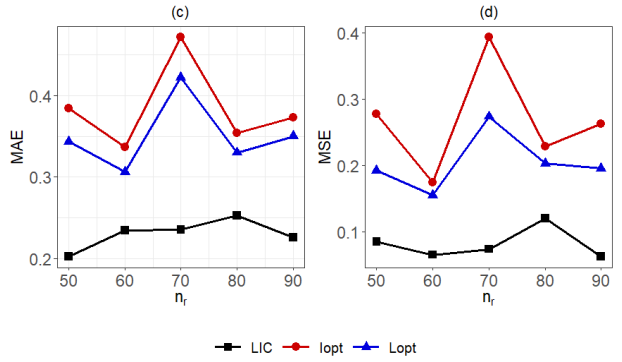

Fig. 10. The values of MAE and MSE with $n_r = (50, 60, 70, 80, 90)$ and $n = 3000$.

Specifically, as K increases, the error evaluation indicators MAE and MSE initially exhibit a significant decreasing trend, followed by an increasing trend. Particularly noteworthy is that when K reaches 20, both MAE and MSE reach their minimum values, specifically 0.0448467 and 0.00775919, respectively, at which point the performance of LIC reaches its optimal state.

Disregarding the effects of endpoint influences and examining the variation of the parameter n_r, it is observed that the error values undergo corresponding changes. Specifically, when n_r equals 70, the error values reach a minimum point. These detailed observations and analyses indicate that changes in both the number of partition blocks K and the parameter n_r significantly impact the overall performance of the model.

Case 3. $X_2 = (X_{2ij}) \in \mathbb{R}^{n_2 \times p}$, $X_{2ij} \sim Geom(0.6)$ with $(\alpha, \sigma_1, \sigma_2, n_r) = (0.01, 3, 8, 8)$.

• ***Scenario I:*** Setting $K = (5, 10, 15, 20, 25)$ with $n = 6000$.

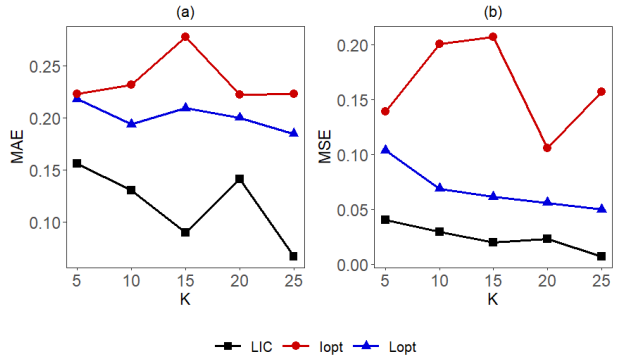

Fig. 11. The values of MAE and MSE with $K = (5, 10, 15, 20, 25)$ and $n = 6000$.

• ***Scenario II:*** Setting $n_r = (50, 60, 70, 80, 90)$ with $n = 3000$.

A thorough analysis of simulation Fig.11 and Fig.12 reveals the significant advantage of LIC criterion over traditional methods in terms of performance.

An observation of Fig.11 indicates that as K increases, the values of MAE and MSE show a decreasing trend overall. Particularly, at $K = 15$, the MAE and MSE under the LIC criterion reach their respective lowest values of 0.09025262 and 0.019957861, marking the optimal state of simulation at this point. Therefore, $K = 15$ can be considered as the ideal number of blocks.

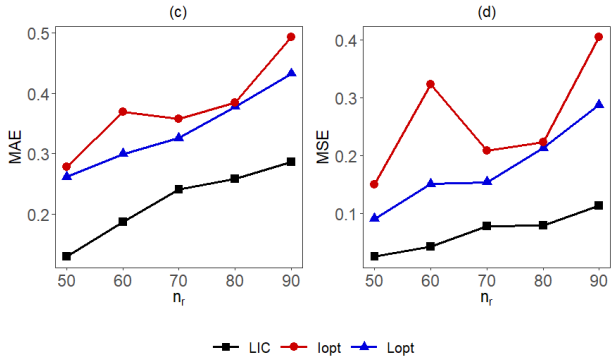

Fig. 12. The values of MAE and MSE with $n_r = (50, 60, 70, 80, 90)$ and $n = 3000$.

D. Simulation Discussion Summary

In conducting a simulation study to delve into the stability and sensitivity, this research examined the performance of this criterion when confronted with redundant error data that follows Uniform, Chi-Square, and Geometric distributions. Research findings indicate that the LIC criterion is highly applicable and performs well with diverse distributions of redundant data. Comparative analyses reveal that changes to the four key parameters $\{n, p, K, n_r\}$ can significantly impact the method's performance. However, despite variations in these parameters, the LIC criterion consistently exhibits lower error rates compared to traditional Iopt and Lopt methods, effectively demonstrating the significant advantages of this method.

This discovery underscores the strength of the LIC criterion and its powerful ability to generalize, making it highly valuable in various scenarios involving redundant distribution data. Looking ahead, it is recommended to explore the ways in which the sensitive parameters $\{n, p, K, n_r\}$ affect model performance. This exploration aims to provide a better understanding of the model's predictive capabilities and offer more precise guidance for model parameter optimization.

V. Conclusion

In the ongoing research, the LIC criterion has been utilized for managing data with specific block lengths and numbers. Future efforts will strive to transcend the limitation of block numbers being constrained by the total number of data points (n), thereby enhancing flexibility in handling diverse data distributions and validating the effectiveness of the method. Additionally, the methodology is intended for application in other statistical models, such as probabilistic, factor, and high-dimensional models, prevalent in fields like social sciences, bioinformatics, and financial data analysis.

Integration of these models with the proposed method promises more efficient management of redundant data, facilitating more precise and reliable statistical analysis outcomes. In multi-modal data contexts, eliminating redundant information becomes crucial, and the redundant data elimination technique has been successfully implemented within the adaptive mixed model framework.

Furthermore, considerations are being given to extending the treatment of error terms by incorporating Laplace White Noise and Symmetric Ergodic Noise variants into the Gaussian normal distribution's ε, thereby enhancing the versatility and practicality of the method for diverse application scenarios.

In conclusion, the objective is to develop a flexible and efficient methodology for managing complex data distributions and models. This method holds significant potential to contribute to data analysis and drive advancements in related fields through continued research and exploration.

Data availability

Our primary focus is on evaluating the performance of the proposed LIC criterion in analyzing simulated data by the LIC package. URL: https://CRAN.R-project.org/package=LIC.

References

[1] G. Guo, Y. Sun, and X. Jiang, "A partitioned quasi-likelihood for distributed statistical inference," *Computational Statistics*, vol. 35, pp.1577–1596, 2020.

[2] G. Guo, Y. Sun, G. Qian, and Q. Wang, "LIC criterion for optimal subset selection in distributed interval estimation," *Journal of Applied Statistics*, 2023, 50(9): 1900-1920.

[3] Q. Wang, G. Guo, G. Qian, and X. Jiang, "Distributed online expectation-maximization algorithm for Poisson mixture model," *Applied Mathematical Modelling*, 2023, 124(2023): 734–748.

[4] Y. Li, 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.

[5] L. Song, G. Guo, "Full Information Multiple Imputation for Linear Regression Model with Missing Response Variable," *IAENG International Journal of Applied Mathematics*, vol 54, pp.77-81,2024.

[6] G. Guo, C. Wei, and G. Q. Qian, "Sparse online principal component analysis for parameter estimation in factor model," *Computational Statistics*, vol. 38, no. 2, pp. 1095-1116. 2022.

[7] C. Huang, X. Huo, "A distributed one-step estimator," *Mathematical Programming*, 2019, 174(1), 41-76.

[8] J. Wang, M. Kolar, N. Srerbo, "Distributed multi-task learning," *In Artificial Intelligence and Statistics*, 2016, 751-760.

[9] C. Qian, G. Li, C. Feng, "Distributed Pareto Optimization for Subset Selection," *In IJCAI*, 2018, 1492-1498.

[10] G. Guo and L. Lu, "Parallel Bootstrap and Optimal Subsample Lengths in Smooth Function Models," *Communications in Statistics–Simulation and Computation*, vol. 45, pp. 2208-2231, 2016.

[11] H. Battey, J. Q. Fan, H. Liu, "Distributed Estimation and Inference with Statistical Guarantees," *ArXiv preprint arXiv*, 2015, 1509.05457.

[12] S. Minsker, N. Strawn, "Distributed Statistical Estimation and Rates of Convergence in Normal Approximation," *Electronic Journal of Statistics*, 2017, 13(2):5213-5252.

[13] B. Mirzasoleiman, A. Karbasi, R. Sarkar, "Distributed Submodular Maximization," *Journal of Machine Learning Research*, 2014, 17(1):8330-8373.

[14] Y. Zhang, J. C. Duchi, M. J. Wainwright, "Divide and Conquer Kernel Ridge Regression: A Distributed Algorithm with Minimax Optimal Rates," *Journal of Machine Learning Research*, 2013, 30(1):592-617.

[15] Q. Cheng, H. Wang, M. Yang, "Information-based optimal subdata selection for big data logistic regression," *Journal of the American Statistical Association*, 2019, 114(525):393-405.

[16] G. Guo, R. Niu, G. Qian, and T. Lu, "Trimmed scores regression for k-means clustering data with high-missing ratio," *Communications in Statistics - Simulation and Computation*, vol. 53, pp. 2805-2821.

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

[18] G. Guo, H. Song, and L. Zhu, "The COR criterion for optimal subset selection in distributed estimation," *Statistics and Computing*, vol. 34, pp. 163-176.

[19] G. Guo, W. Shao, L. Lin, and X. Zhu, "Parallel Tempering for Dynamic Generalized Linear Models," *Commun.Statist.-Theory Meth.*, vol. 45, pp. 6299-6310, 2016.

[20] G. Guo, W. You, L. Lin, and G. Qian, "Covariance Matrix and Transfer Function of Dynamic Generalized Linear Models," *Journal of Computational and Applied Mathematics*, vol. 296, pp. 613–624, 2016.

[21] G. Guo, G. Qian, L. Lin, and W. Shao, "Parallel inference for big data with the group bayesian method," *Metrika*, vol. 84, pp. 225-243, 2021.

[22] G. Guo, G. Qian, and L. Zhu, "A scalable quasi-Newton estimation algorithm for dynamic generalized linear model," *Journal of Nonparametric Statistics*, vol. 34, pp. 917-939, 2022.