Distributed Estimation of Redundant Data

Qianwen Liu, Guangbao Guo

***Abstract*—In the era of big data, traditional statistical estimation methods are challenged by expanding data scales. Distributed estimation methods arise to address this, enhancing efficiency by processing data across multiple nodes. However, redundant data handling is crucial in distributed estimation, as it can compromise accuracy and increase variance. The LIC-based distributed estimation method mitigates this by generating random point sets, improving accuracy. Distributed estimation methods hold significant promise in managing redundant data, gaining increasing attention and research. This article delves into optimal subset selection for LIC-based distributed estimation with redundant data.**

***Index Terms*—redundant data, distributed estimation, big data, LIC criterion.**

I. Introduction

In recent years, distributed estimation methods have made significant progress in processing large-scale data sets, and their applications in multiple fields have also received widespread attention. In terms of dealing with redundant data, distributed estimation methods have significant advantages. Firstly, by allocating data to multiple nodes for processing, the processing speed and efficiency can be greatly enhanced. Secondly, distributed estimation methods are more effective in handling noise and outliers in large-scale datasets, thereby enhancing the accuracy and reliability of the estimation. Moreover, distributed estimation methods can effectively integrate data from different sources, improving the quality and availability of data. In statistics, the optimal subset method is a process of finding the best predictive variables that can maximize the predictive ability of the prediction model. However, traditional optimal subset methods encounter computational complexity and slow speed when dealing with high-dimensional data. To address these issues, we propose an optimal subset method based on LIC distributed estimation. This method combines the advantages of LIC distributed estimation to effectively process high-dimensional data while enhancing computational efficiency and accuracy.

A. Current Research Status

Recently, the development of distributed statistical inference and high-dimensional statistical methods has garnered significant attention. LIC proposed by Guo et al. [1] has demonstrated the performance in optimal subset selection for distributed interval estimation. Building on this foundation, see J. Lederer [2] for related work. See also Guo et al. [3], Li et al. [4], Guo et al. [5] - [7], Wang et al. [8], Song et al. [9], Guo et al. [10] - [12]. Additionally, the partitioned quasi-likelihood method developed by Guo et al. [13] provides a novel perspective for distributed statistical inference, emphasizing the significance of data partitioning in the analytical process. These studies have not only advanced the development of statistical computing but also provided theoretical support and technical assurance for practical applications.

B. Our Work

This article will explain the basic principles, implementation steps, and advantages of the optimal subset method using LIC distributed estimation. The goal is to provide readers with a deeper understanding of the method and help them grasp its advantages and effectiveness in handling high-dimensional data. Through the introduction and research presented in this article, we hope to offer readers a novel perspective and approach to solving the challenges faced by traditional optimal subset methods when dealing with high-dimensional data. For distributed estimation of redundant data, we need to research and develop effective algorithms and methods. This may involve multiple aspects, such as data preprocessing, feature selection, model selection, and so on, with the goal of improving the accuracy and efficiency of estimation. At the same time, we need to explore the advantages and disadvantages of different methods and their application scopes, in order to make reasonable choices and optimizations in practical applications. To ascertain the efficacy of the proposed approach, we need to conduct a sufficient experimental design and its implementation. This includes selecting appropriate datasets, designing reasonable experimental plans, controlling experimental conditions, collecting, and analyzing experimental results, etc. Through experiments, we can compare the performance of different methods in dealing with redundant data and further optimize and improve the method.

II. Theorem

A. Notation

The vector $u_{I_{opt}}$ represents a sub-residual vector. The original identity matrix is denoted as $I_{n_{I_{opt}} \times n_{I_{opt}}}$. The regression coefficient vector β consists of β_1 through β_p, expressed as $\beta = (\beta_1, ..., \beta_p)^\top$. The unknown variance is denoted by σ^2.

The optimal estimate of β based on the sub-matrix $X_{I_{opt}}$ is given by $\widehat{\beta}_{I_{opt}} = X^+_{I_{opt}} y_{I_{opt}}$ where $X^+_{I_{opt}}$ is the pseudoinverse of $X_{I_{opt}}$, calculated as $VD^+U^\top$ using singular value decomposition.

Manuscript received January 18, 2024; revised December 18, 2024.

This work was supported by the National Social Science Foundation Project under project ID 23BTJ059, a grant from Natural Science Foundation of Shandong under project ID ZR2020MA022, and the National Statistical Research Program under project ID 2022LY016.

Qianwen Liu is a postgraduate student of Mathematics and Statistics, Shandong University of Technology, Zibo, China. (e-mail: Qianwen001215@163.com).

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

B. Theorem and proof

The focus is on distributed linear regression models: $Y_{I_{opt}} = X_{I_{opt}}\beta + u_{I_{opt}}$, $u_{I_{opt}} \sim N\left(0, \sigma^2 I_{n_{I_{opt}} \times n_{I_{opt}}}\right)$, $k = 1, ..., K_n$.

The unique least-squares estimator is given by $\widehat{\beta}_{I_{opt}} = (X_{I_{opt}}^\top X_{I_{opt}})^{-1} X_{I_{opt}}^\top Y_{I_{opt}}$ is orthogonal square matrices. $U \in \mathbb{R}^{n_{Iopt} \times n_{Iopt}}$ and $V \in \mathbb{R}^{p \times p}$ are orthogonal square matrices. The diagonal matrix $D \in \mathbb{R}^{n_{Iopt} \times p}$ is (with $D_{ij} = 0$ for $i \neq j$) such that $X_{I_{opt}} = UDV^\top$. The matrix $X_{I_{opt}}$ is a $n_{I_{opt}} \times p$ sub-matrix of X with $n_{I_{opt}} \geq p$.

$$\begin{aligned}(X^\top X)^{-1}X^\top &= ((UDV^\top)^\top(UDV^\top))^{-1}(UDV^\top)^\top \\ &= (VDU^\top)(UDV^\top))^{-1}(VDU^\top) \\ &= (V(D^+)^2V^\top)^{-1}(VDU^\top) \\ &= ((V^\top)^{-1}(D^+)^2V^{-1})(VDU^\top) \\ &= VD^+U^\top.\end{aligned}$$

Theorem 1. For any least-squares solution $\widehat{\beta}_{I_{opt}}$, it is fundamental to note that the squared Euclidean norm of the residual vector resulting from the prediction, $X_{I_{opt}}\beta_{I_{opt}} - X_{I_{opt}}\widehat{\beta}_{I_{opt}}$, is equivalent to the squared norm of a transformed noise vector. Specifically,

$$\|X_{I_{opt}}\beta_{I_{opt}} - X_{I_{opt}}\widehat{\beta}_{I_{opt}}\|_2^2 = \|UDD^+U^\top u_{I_{opt}}\|_2^2,$$

where the equality stems from the mathematical properties of the least-squares estimator and the singular value decomposition (SVD) of the relevant design matrix $X_{I_{opt}}$. Here, U, D, and D^+ represent the unitary matrix, diagonal matrix of singular values, and pseudoinverse of the diagonal matrix, respectively, from the SVD of $X_{I_{opt}}$, and $u_{I_{opt}}$ is the relevant subset of the noise vector u.

Assuming that the noise vector $u \sim \mathcal{N}(0, \sigma^2 I_{n\times n})$ where $\sigma \in (0, \infty)$ denotes the standard deviation, we can derive a risk bound for the average squared prediction error. This risk bound quantifies the expected deviation of the prediction error, normalized by the sample size n, and is given by

$$E\left[\frac{\|X_{I_{opt}}\beta_{I_{opt}} - X_{I_{opt}}\widehat{\beta}_{I_{opt}}\|_2^2}{n}\right] = \frac{\sigma^2 rank\left[X_{I_{opt}}\right]}{n}.$$

This equation elegantly relates the expected prediction error to the standard deviation of the noise, the rank of matrix $X_{I_{opt}}$ and the sample size n. The higher the noise level or model complexity, the greater the expected prediction error, while a larger sample size typically helps reduce the prediction error.

Proof.

$$\begin{aligned}&\|X_{I_{opt}}\beta_{I_{opt}} - X_{I_{opt}}\widehat{\beta}_{I_{opt}}\|_2^2 \\ &= \|X_{I_{opt}}\beta_{I_{opt}} - X_{I_{opt}}X_{I_{opt}}^+ y_{I_{opt}}\|_2^2 \qquad \text{our choice of } \widehat{\beta}_{I_{opt}} \\ &= \|X_{I_{opt}}\beta_{I_{opt}} - X_{I_{opt}}X_{I_{opt}}^+ \left(X_{I_{opt}}\beta_{I_{opt}} + u_{I_{opt}}\right)\|_2^2 \\ &\qquad \text{model assumptions: } y_{\mathrm{I_{opt}}} = X_{I_{opt}}\beta + u_{I_{opt}} \\ &= \|X_{I_{opt}}X_{I_{opt}}^+ u_{I_{opt}}\|_2^2 \qquad X_{I_{opt}}^+ X_{I_{opt}} = I_{n_{I_{opt}}\times n_{I_{opt}}} \\ &= \|UDV^\top VD^+U^\top u_{I_{opt}}\|_2^2 \qquad SVD \\ &= \|UDD^+U^\top u_{I_{opt}}\|_2^2 \qquad V^\top V = E \\ &= \left(UDD^+U^\top u_{I_{opt}}\right)^\top UDD^+U^\top u_{I_{opt}} \\ &= \left(U^\top u_{I_{opt}}\right)^\top (DD^+)^\top U^\top UDD^+U^\top u_{I_{opt}} \\ &= \left(U^\top u_{I_{opt}}\right)^\top (DD^+)^\top DD^+U^\top u_{I_{opt}} \qquad Uorthogonal \\ &= \left(U^\top u_{I_{opt}}\right)^\top DD^+DD^+U^\top u_{I_{opt}} \\ &= \left(U^\top u_{I_{opt}}\right)^\top DD^+U^\top u_{I_{opt}}.\end{aligned}$$

Setting $\gamma = \frac{U^\top u_{I_{opt}}}{\sigma}$,

$$\gamma^\top DD^+\gamma \sim \chi^2 rank[X_{I_{opt}}],$$

$$\|X_{I_{opt}}\beta_{I_{opt}} - X_{I_{opt}}\widehat{\beta}_{I_{opt}}\|_2^2 \sim \sigma^2\gamma^\top DD^+\gamma,$$

$$\frac{\|X_{I_{opt}}\beta_{I_{opt}} - X_{I_{opt}}\widehat{\beta}_{I_{opt}}\|_2^2}{n} \sim \frac{\sigma^2\gamma^\top DD^+\gamma}{n},$$

$$E\left[\frac{\|X_{I_{opt}}\beta_{I_{opt}} - X_{I_{opt}}\widehat{\beta}_{I_{opt}}\|_2^2}{n}\right] = \frac{\sigma^2 rank\left[X_{I_{opt}}\right]}{n}.$$

Since U is orthogonal, $\gamma \sim N\left(0, \sigma^2 I_{n_{I_{opt}}\times n_{I_{opt}}}\right)$. Additionally, DD^+ features rank(X) entries of one along its main diagonal, with all other entries being zero. These insights lead to the conclusion that the quadratic form $\gamma^\top DD^+\gamma \sim \chi^2_{rank(X)}$. The result follows from the property that the mean of a Chi-Squared distribution is equal to its degrees of freedom. □

III. Simulation

A. Simulation preparation

The (X, Y) is from the model $Y_i = X_i\beta + \varepsilon_i$, $\varepsilon_i \sim N\left(0, \sigma_i^2 I_{n\times n}\right)$ for $i = 1, 2$. It is known that X is composed of (X_1, X_2) and Y is composed of (Y_1, Y_2).

$$\begin{aligned}X_1 &= (X_{1ij}) \in \mathbb{R}^{n_1\times p},\ X_{1ij} \sim N(0, 4); \\ X_2 &= (X_{2ij}) \in \mathbb{R}^{n_2\times p},\ X_{2ij} \sim F(X); \\ Y_1 &= X_1\beta + \varepsilon_1,\ n_1 = [1, \ldots, (n - n_r)]; \\ Y_2 &= X_2\beta + \varepsilon_2,\ n_2 = [1, \ldots, n_r].\end{aligned}$$

Additionally, it is known that $\beta \sim Unif(0.5, 2)$, and $\varepsilon \sim (\varepsilon_1, \varepsilon_2)$, where $\varepsilon_1 \sim N(0, 8)$ and $\varepsilon_2 \sim N(0, 20)$.

The purpose of this section is to evaluate the predictive accuracy of three different models (LIC, Lopt, Iopt) with varying (n, p, K, n_r). To assess the predictive accuracy of the models in a data simulation environment, we use metrics such as the MSE and MAE to quantify the deviation between the true values and the predicted values.The MSE and MAE, which are used to assess prediction error, are defined as follows:

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

Among them, n is sample sizes, K is the number of subsets that data is divided into, p is feature dimensions, α is the significance level, σ_1 and σ_2 are the standard deviation of noise, and n_r is a partitioning point of the data set, used to distinguish between two different subsets of data.

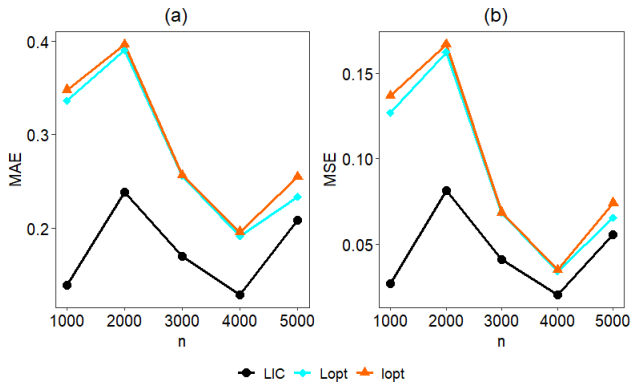

Fig. 1. The values of MAE and MSE with $p = 8$.

B. Case 1

$$X_2 = (X_{ij}) \in \mathbb{R}^{n_2 \times p},\ X_{2ij} \sim F(n, 1).$$

Scenario 1: Setting $(K, \alpha, n_r, p) = (10, 0.05, 50, 8)$; $n = (1000, 2000, 3000, 4000, 5000)$.

As depicted in the Fig. 1, with fixed (p, K, n_r), the impact of varying n on the results is studied. The trends of the curves are roughly the same, with the LIC being lower compared to Iopt and Lopt. As the value of the variable increases, both MAE and MSE values first rise and then fall. It can also be seen that, compared to Iopt and Lopt, LIC has lower MAE and MSE values. As the variable value reaches 4000, the model fitting is optimal, with MAE and MSE values of 0.1286 and 0.02019. Therefore, the optimal value for n is 4000, which is the best value for model fitting.

Scenario 2: Setting $(K, \alpha, n_r, n) = (10, 0.05, 50, 1000)$; $p = (8, 9, 10, 11, 12)$.

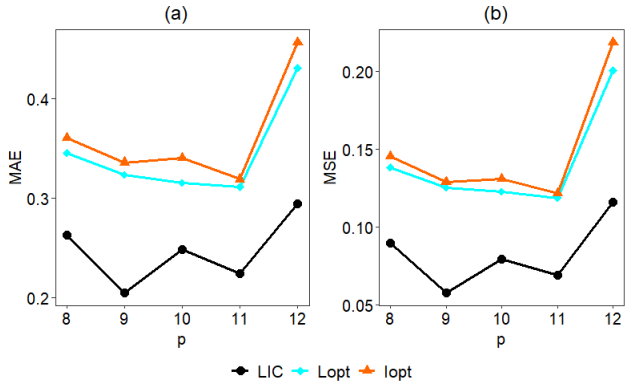

Fig. 2. The values of MAE and MSE with $n = 1000$.

As shown in Fig. 2, with fixed (n, K, n_r), the MAE and MSE change as the value of p varies and the trends of the curves are roughly the same. As p increases from 8 to 9, MAE decreases from 0.2629 to 0.2046, and MSE decreases from 0.0895 to 0.0577. Conversely, as p increases from 9 to 10, MAE increases from 0.2046 to 0.2479, and MSE increases from 0.0577 to 0.0794. It can be concluded that the model fitting is optimal as p is set to 9.

Scenario 3: Setting $(\alpha, n_r, n, p) = (0.05, 50, 3000, 8)$; $K = (4, 5, 6, 8, 10)$.

As depicted in the Fig. 3, with fixed (n, p, n_r), the MAE and MSE curves exhibit trends that closely follow the changes in K. Initially, as K increases, both MAE and MSE decrease. However, once K reaches a value of 6, both error metrics start to increase. At this point, with K set to 6, the MAE and MSE achieve their lowest values of 0.1591 and 0.0291, respectively, indicating the optimal fitting state for the model.

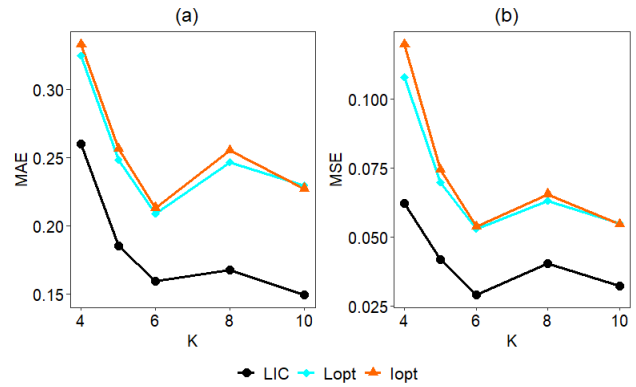

Fig. 3. The values of MAE and MSE with $n_r = 50$.

Scenario 4: Setting $(K, \alpha, n, p) = (10, 0.05, 3000, 8)$; $n_r = (30, 40, 50, 60, 70)$.

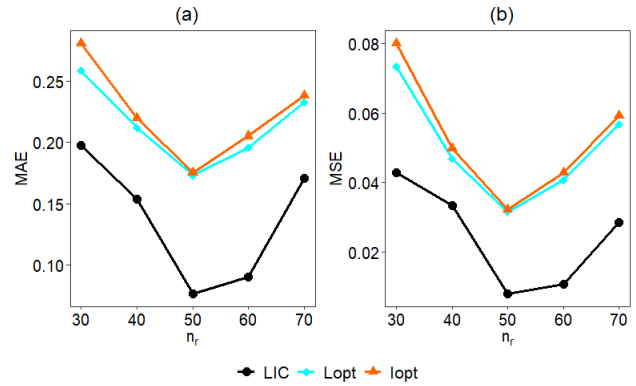

Fig. 4. The values of MAE and MSE with $K = 10$.

As shown in Fig. 4, with fixed (n, p, K), the trends of the MAE and MSE curves are found to be consistent with the variation of n_r. As n_r increases, both MAE and MSE values gradually decrease. Specifically, as n_r is between 50 and 60, MAE range from 0.0768 to 0.0906, and MSE range from 0.0079 to 0.0107. Based on this observation, it is concluded that the optimal fitting effect is achieved as n_r is set to 50.

This study investigates the data simulation and model fitting process using the F-distribution and the LIC criterion. Comparing Scenario 1 and Scenario 2 shows that the trends of MAE and MSE are similar, and compared to the Lopt and Iopt , the LIC criterion performs better, with smaller MAE and MSE values, indicating its superior stability. The detailed numerical analysis is provided in Scenario 1 and Scenario 2. Scenario 3 and Scenario 4 shows that as the number of blocks K, and the parameter n_r change, the MAE and MSE values first decrease and then increase. This indicates that the performance of the LIC criterion improves initially and then starts to deteriorate. Therefore, the overall best performance is achieved as $K = 6$ and $n_r = 50$. The detailed numerical analysis is provided in Scenario 3 and Scenario 4.

C. Case 2

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

Scenario 1: Setting $(K, \alpha, n_r, p) = (10, 0.05, 50, 8)$; $n = (1000, 2000, 3000, 4000, 5000)$.

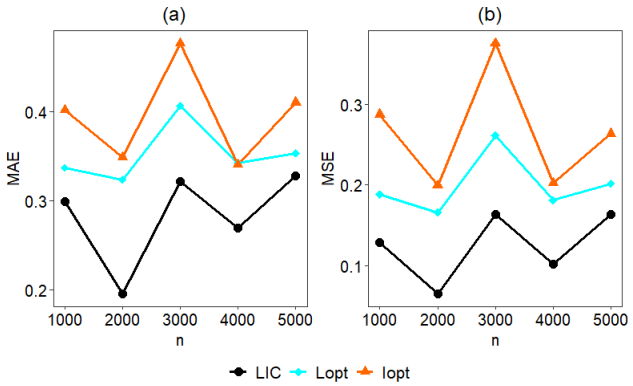

Fig. 5. The values of MAE and MSE with $p = 8$.

As depicted in the Fig. 5, with fixed (p, K, n_r), the impact of varying n on the results is studied. The trends of the curves are roughly the same, with the LIC being lower compared to Iopt and Lopt. As n increases, both MAE and MSE values initially rise before subsequently declining. The model achieves its best fit as n is 2000, at which point the MAE and MSE values are 0.1954 and 0.0647. Consequently, the optimal value for n in terms of model fitting is 2000.

Scenario 2: Setting $(K, \alpha, n_r, n) = (5, 0.05, 50, 1000)$; $p = (8, 9, 10, 11, 12)$.

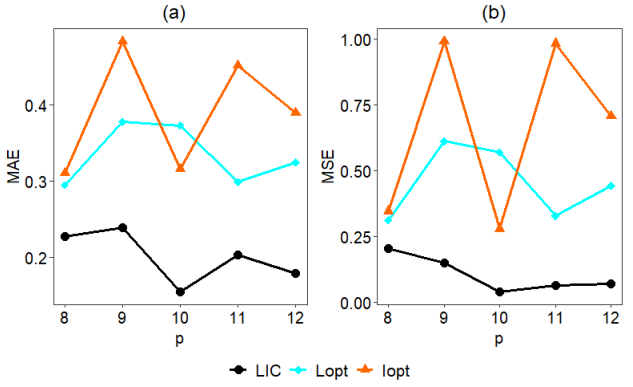

Fig. 6. The values of MAE and MSE with $n = 1000$.

As shown in Fig. 6, with fixed (n, K, n_r), changes in p affect both the MAE and MSE, with the trends of the curves being roughly the same and the LIC being lower compared to Iopt and Lopt. As p increases from 9 to 10, MAE decreases from 0.2386 to 0.1549, and MSE decreases from 0.1509 to 0.0401. However, as p further increases from 10 to 11, MAE increases to 0.2036, and MSE increases to 0.0644. Based on these observations, it is concluded that the optimal fitting effect is achieved as p is set to 9.

Scenario 3: Setting $(\alpha, n_r, n, p) = (0.05, 50, 3000, 8)$; $K = (4, 5, 6, 8, 10)$.

As depicted in the Fig. 7, with fixed (n, p, n_r), the MAE and MSE curves exhibit trends that closely follow the changes in K, and the LIC being lower compared to Iopt and Lopt. Initially, as K increases, both MAE and MSE

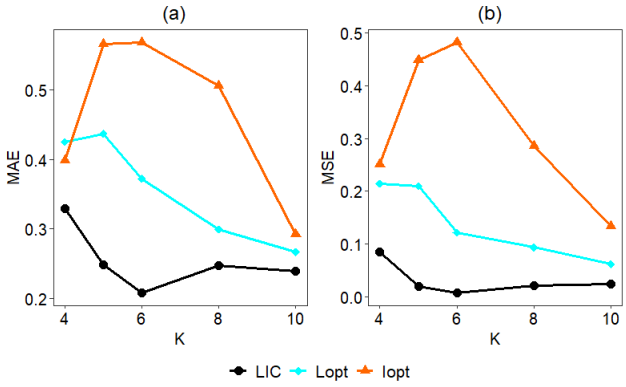

Fig. 7. The values of MAE and MSE with $n_r = 50$.

decrease. However, they begin to increase once K reaches a value of 6. At this point, with K set to 6, the MAE and MSE values are 0.2073 and 0.0065, indicating the model has achieved its best fit.

Scenario 4: Setting $(K, \alpha, n, p) = (10, 0.05, 3000, 8)$; $n_r = (30, 40, 50, 60, 70)$.

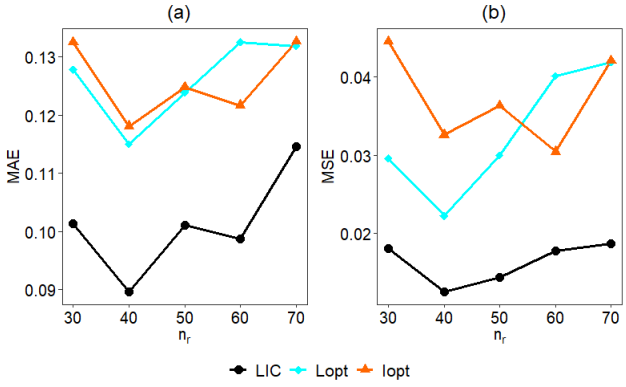

Fig. 8. The values of MAE and MSE with $K = 10$.

As shown in Fig. 8, with fixed (n, p, K), the trends of the MAE and MSE curves closely mirror the changes in n_r. As n_r increases, both MAE and MSE values gradually decrease. Specifically, as n_r is between 30 and 40, MAE range from 0.1013 to 0.0896, and MSE range from 0.0181 to 0.0125. Therefore, it is concluded that the optimal fitting effect is achieved as n_r is 40.

The study observed changes in the trends of MAE and MSE, which are crucial metrics for assessing the quality of model fit. By systematically modifying the model parameters, the research revealed distinct patterns of model fitting performance under different conditions. Secondly, the simulation results indicate that the LIC criterion generally outperforms both Iopt and Lopt in terms of achieving lower MAE and MSE values. This demonstrates the excellent stability of the LIC criterion, leading to a better fit and more accurate predictions. This advantage of LIC is particularly notable in the presence of redundant data, where other methods may struggle to differentiate between relevant and irrelevant information. The robustness of LIC in handling redundant data can be attributed to its ability to incorporate additional constraints or regularization techniques that help in selecting the most informative variables while minimizing overfitting. This is crucial in statistical modeling, as redun-

dant data can lead to increased complexity and decreased interpretability of the model.

D. Case 3

$$X_2 = (X_{ij}) \in \mathbb{R}^{n_2 \times p},\ X_{2ij} \sim weibull(3,1).$$

Scenario 1: Setting $(K, \alpha, n_r, p) = (10, 0.05, 50, 8)$; $n = (1000, 2000, 3000, 4000, 5000)$.

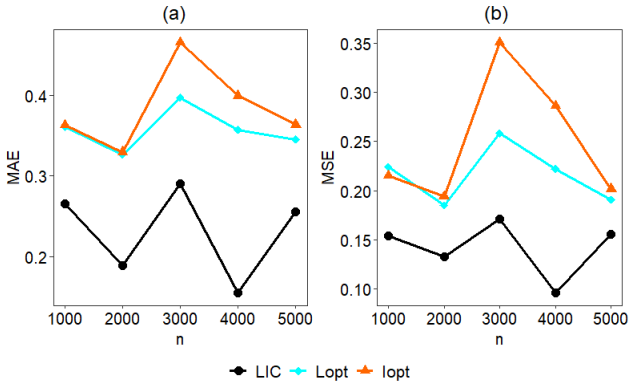

Fig. 9. The values of MAE and MSE with $p = 8$.

As depicted in the Fig. 9, with fixed (p, K, n_r), the impact of varying n on the results is studied. The trends of the curves are roughly the same, with the LIC being lower compared to Iopt and Lopt. With an increase in n, both MAE and MSE values initially rise before they start to decline. The result is optimal as n is at 4000, at which point the MAE and MSE values are recorded as 0.1554 and 0.0963, respectively. Consequently, the optimal value for n in terms of model fitting is determined to be 4000.

Scenario 2: Setting $(K, \alpha, n_r, n) = (5, 0.05, 50, 1000)$; $p = (8, 9, 10, 11, 12)$.

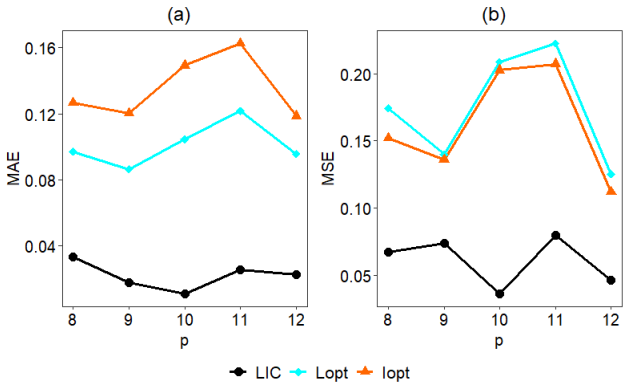

Fig. 10. The values of MAE and MSE with $n = 1000$.

As shown in Fig. 10, with (n, K, n_r) fixed and p varied, the performance of the model fitting is analyzed based on the MAE and MSE. As p increases from 9 to 10, MAE decreases from 0.0179 to 0.0106, and MSE increases from 0.0734 to 0.0359. Conversely, as p increases from 10 to 11, MAE increases to 0.0257, and MSE increases to 0.0793. Based on these observations, it is concluded that the model achieves its best fit as p is at 10.

Scenario 3: Setting $(\alpha, n_r, n, p) = (0.05, 50, 3000, 8)$; $K = (4, 5, 6, 8, 10)$.

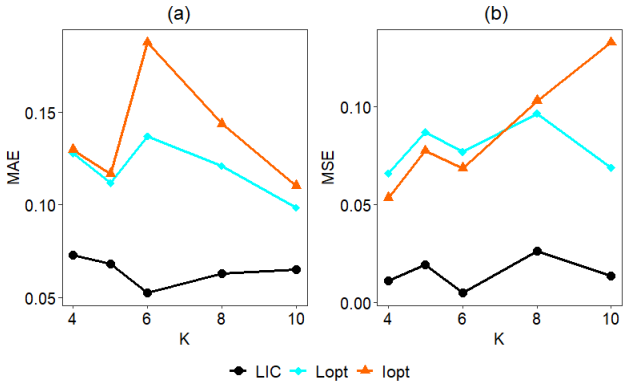

Fig. 11. The values of MAE and MSE with $n_r = 50$.

As depicted in the Fig. 11, with fixed (n, p, n_r), the MAE and MSE curves exhibit trends that closely follow the changes in K. Initially, as K increases, both MAE and MSE values decrease. However, they begin to increase once K reaches the value of 6. At this specific value of $K = 6$, the MAE and MSE values are recorded as 0.0523 and 0.0049, respectively, indicating that the model has achieved its optimal fitting state.

Scenario 4: Setting $(K, \alpha, n, p) = (10, 0.05, 3000, 8)$; $n_r = (30, 40, 50, 60, 70)$.

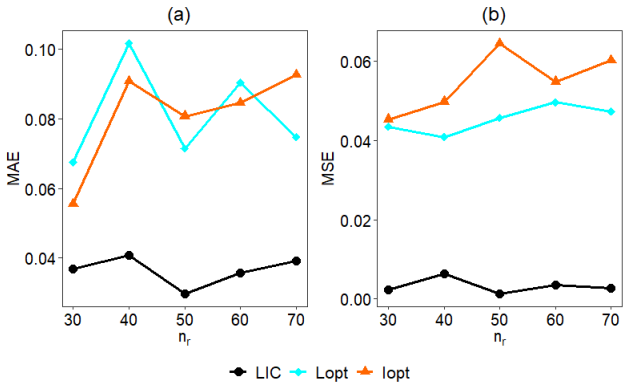

Fig. 12. The values of MAE and MSE with $K = 10$.

As shown in Fig. 12, with fixed (n, p, K), the trends of the MAE and MSE curves are basically the same as changing n_r. As n_r increases, MAE and MSE gradually decrease. As the value of n_r ranges from 40 to 50, MAE and MSE range from 0.0408 to 0.0296 and from 0.0063 to 0.0013, respectively. Therefore, it is concluded that the optimal fitting effect is achieved as n_r is set to 50.

In this study, it can be observed that as n and p change, the model's fitting performance also changes significantly. Increasing the sample size usually helps improve the stability and prediction accuracy of the model, while increasing the number of features may introduce more redundant information, which may affect the interpretability and generalization ability of the model. The presence of redundant data can significantly reduce the predictive accuracy and interpretability of models. This is because redundant data increases the complexity of the model, making it more susceptible to noise during the fitting process, resulting in unstable prediction results. The LIC criterion balances the fitting effect and generalization ability of the model by

considering the complexity of the model and the amount of information in the data, thus helping to avoid overfitting and underfitting problems. The experimental results show that the LIC criterion performs well in model selection, and can select models that have both good fitting effects and good generalization capabilities.

IV. Conclusion

The current study has delved into the realm of distributed estimation for redundant data, offering insights into the challenges and opportunities presented by this domain. Through the exploration of block strategies, we have gained a deeper understanding of how flexibility and optimization can play a crucial role in handling large-scale datasets with redundant information.

The paper emphasizes the necessity of devising adaptive block strategies that are not merely sensitive to the scale of the data, but also cognizant of the specific attributes of redundant data. The efficacy of such strategies crucially depends on the capacity to dynamically adjust block length and quantity according to actual requirements, thereby maximizing both the efficiency and accuracy of distributed estimation. Furthermore, the study underscores the potential to broaden our scope to encompass redundant data exhibiting diverse distributional properties.

In particular, the application of time series models to distributed estimation emerges as a promising avenue. The inherent redundancies within time series data present unparalleled opportunities for enhancing estimation accuracy and efficiency, particularly when synergized with suitable block strategies. Consequently, we advocate for intensified research in this domain to capitalize fully on the benefits that time series characteristics can offer within the context of distributed estimation.

In conclusion, the study underscores the significance of distributed estimation in handling redundant data and highlights the need for continued innovation in block strategies, adaptability to different data distributions, and exploration of time series models.

V. Further Work

In future research, for the problem of redundant data, we can further explore the flexibility and optimization of block strategies. In large-scale datasets, the length and quantity of blocks may need to be adjusted according to actual needs. Future research can explore how to design more flexible block strategies to better adapt to different types of data and problems.

Regarding different forms of distributed redundant data, current research primarily focuses on those with specific distributions, like the Gaussian distribution. Future studies can delve into managing redundant data adhering to other distributions, such as the Poisson and exponential distributions, among others. This endeavor will broaden the application spectrum of distributed estimation techniques. As it comes to applying time series models, a notable quantity of redundant information is frequently observed in time series data. Exploring the utilization of time series models for distributed estimation represents a promising avenue. By leveraging the inherent characteristics of time series, we may enhance both the accuracy and efficiency of distributed estimation.

Data Availability

We utilized the LIC criterion to fit the data matrices of three distributions: F-distribution, T-distribution, and Weibull distribution, thereby simulating redundant data. This simulation method is used to study the application of the LIC criterion in redundant data distributed estimation. The implemented LIC criterion has been integrated into an R package. URL: https://CRAN.Rproject.org/package=LIC.

References

[1] G. Guo, Y. Sun, G. Qian, and Q. Wang, "LIC criterion for optimal subset selection in distributed interval estimation," *Journal of Applied Statistics*, vol. 50, no. 9, pp. 1900-1920, 2022.

[2] J. Lederer. "Fundamentals of High-Dimension Statistics," *Switzerland. Springer Nature Switzerland*, AG. 2020. 1.

[3] G. Guo, W. You, G. Qian, and W. Shao, "Parallel maximum likelihood estimator for multiple linear regression models," *Journal of Computational and Applied Mathematics*, vol. 273, pp. 251-263, 2015.

[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, pp. 205-211, 2024.

[5] G. Guo, "Parallel statistical computing for statistical inference," *Journal of Statistical Theory and Practice*, vol. 6, no. 3, pp. 536-565, 2012.

[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] 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.

[8] Q. Wang, G. B. Guo, G. Q. Qian, and X. J. Jiang, "Distributed online expectation-maximization algorithm for Poisson mixture model," *Applied Mathematical Modelling*, vol. 124, pp. 734-748, 2023.

[9] L. Song, G. Guo, "Full Information Multiple Imputation for Linear Regression Model with Missing Response Variable," *IAENG International Journal of Applied Mathematics*, vol. 54, no. 1, pp. 77-81. 2024.

[10] 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, 2024.

[11] G. Guo, M. Yu, and G. Qian, "ORKM: Online regularized K-means clustering for online multi-view data," *Information Sciences*, vol. 680, Article ID 121133, 2023.

[12] 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, 2023.

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