The Optimal Subset Estimation of Distributed Redundant Data

Congfan Zhang, Guangbao Guo

***Abstract*—This paper discusses the issue of estimating the optimal subset from redundant distributed data in the context of the big data environment. A method based on LIC is proposed, which can effectively extract useful information from distributed redundant data and identify the optimal data subset. Through a series of experiments, the paper verifies the performance of the LIC method in improving data quality and utilization efficiency, and evaluates its effectiveness. The results show that the LIC method has significant advantages in handling large-scale, high-dimensional and complex distributed datasets.**

***Index Terms*—distributed redundant data, optimal subset estimation, LIC method, performance evaluation.**

I. Introduction

WITH the continuous expansion of data scale, distributed data storage and processing have become essential means for big data analysis. However, this data processing method also poses the challenge of data redundancy. Redundancy refers to the presence of duplicate or similar information within the data, while distribution signifies that the data is scattered across multiple sources or databases. Redundant data not only occupies additional storage space but also potentially leads to the waste of computing resources and may even compromise the accuracy of data analysis. Therefore, accurately estimating the optimal subset of redundant distributed data is of significant practical importance. This article employs LIC technology to analyze the data in order to extract useful information from redundant distributed data. LIC technology can comprehensively consider the temporal and spatial distribution characteristics of the data, thus enabling a more effective extraction of redundant information.

A. Current Research Status

The development of redundant distributed data processing has attracted significant attention. In the field of distributed statistical inference, Guo et al. [7] introduced the LIC criterion in their study, emphasizing its effectiveness in enhancing interval estimation within distributed systems. The LIC criterion serves as a powerful methodology for identifying the most informative subsets of data, which is particularly important in scenarios with limited computational resources and data distributed across multiple locations. Guo [2] conducted a comprehensive exploration of parallel statistical computation, arguing that this approach provides a transformative pathway for enhancing statistical inference. In his paper titled “Parallel Statistical Computation and Statistical Inference,” Guo discusses the growing demand for efficient computational techniques in the context of increasingly large and complex datasets.

The references Wang et al. [1], Guo et al. [3], Guo et al. [4], Li et al. [5] and Guo et al. [8] serve as major sources for other key studies on distributed statistical inference. Reference J et al. [6] emphasizes research on high-dimensional statistics, while references Guo et al. [9] and Guo et al. [10] highlight studies on online learning algorithms. References Song et al. [11] focus on inference related to missing response variables. References Guo et al. [12]-[14] explore knowledge related to parallel inference and dynamic linear models. Overall, the literature reflects a concerted effort to enhance statistical methodologies through distributed computing, thereby paving the way for more efficient and effective analyses of large-scale data.

B. Our Work

This article presents a practical and efficient method for solving the problem of optimal subset estimation in redundant distributed data. Firstly, we introduce the principles and implementation details of the LIC criterion, as well as its advantages in extracting redundant information. Secondly, we present experimental results and performance evaluations to verify the feasibility and superiority of this method through actual data. Finally, we explore the application prospects and future development directions of this method in related fields. To validate the effectiveness of the method presented in this article, a comprehensive set of experiments was conducted. Firstly, we constructed a simulation data set that simulates redundant distributed data of varying sizes, dimensions, and distributions. Then, we applied the method we proposed to process the simulation data and compared its performance with existing methods. The experimental results demonstrate that LIC exhibits significant advantages in handling large-scale, high-dimensional, and complex distributed data sets.

II. Method and Theorem

A. Notation

We define the the empirical covariance matrix $A = X_{I_{opt}} X_{I_{opt}}^T \setminus n_{I_{opt}}$ where $X_{I_{opt}} \sim N_{n_{I_{opt}}}\left[0_{n_{I_{opt}}}, \Sigma_{I_{opt}}\right]$, I_{opt} is optimal subsets indicate set function, $\Sigma_{I_{\mathrm{opt}}}$ is symmetric and invertible matrix, $X^1, \ldots, X^{n_{I_{opt}}}$ is independent realizations, in addition $0_{n_{I_{opt}}}$ is Gauss distribution mean

Manuscript received September 3, 2024; revised December 19, 2024.

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

Congfan Zhang is a postgraduate student of Mathematics and Statistics, Shandong University of Technology, Zibo, China. (e-mail:Congfan0706@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).

value matrix, $\mathbb{R}^{n_{I_{opt}}}$is a set of real numbers, $a \in \mathbb{R}^{n_{\mathrm{I_{opt}}}}$is a matrix on the set of $n_{I_{opt}}$ dimensional real numbers.

B. Theorem and Proof

Theorem 1. The probability that the matrix $X_{\mathrm{I_{opt}}} X_{\mathrm{I_{opt}}}^T / n_{\mathrm{I_{opt}}}$ belongs to the set S_d^+ of symmetric and positive definite matrices, where $S_d^+ = \{AA^T \mid A \in \mathbb{R}^{d \times p}\}$ and A is a real-valued matrix, is unity. In other words, it holds with certainty that $X_{\mathrm{I_{opt}}} X_{\mathrm{I_{opt}}}^T / n_{\mathrm{I_{opt}}}$ is a symmetric and positive definite matrix.

Proof of Theorem 1.

In this section, we capitalize on the fundamental principle that sets residing in subspaces of dimensions lesser than their ambient spaces possess a Gauss measure identically equal to zero. This fundamental insight serves as the cornerstone of our subsequent analysis.

The verification of symmetry is rather straightforward. Consider the transpose operation applied to the given expression:

$$\left(\frac{1}{n_{\mathrm{I_{opt}}}} \sum_{i=1}^{n_{\mathrm{I_{opt}}}} X_i X_i^T \right)^T = \frac{1}{n_{\mathrm{I_{opt}}}} \sum_{i=1}^{n_{\mathrm{I_{opt}}}} \left(X_i X_i^T \right)^T$$

Given that the transpose of a matrix product AB is $(AB)^T = B^T A^T$, and noting that $X_i^T X_i$ is symmetric (i.e., $(X_i^T X_i)^T = X_i^T X_i$), we have:

$$\frac{1}{n_{\mathrm{I_{opt}}}} \sum_{i=1}^{n_{\mathrm{I_{opt}}}} \left(X_i X_i^T \right)^T = \frac{1}{n_{\mathrm{I_{opt}}}} \sum_{i=1}^{n_{\mathrm{I_{opt}}}} X_i X_i^T$$

Thus, the desired symmetry property is established.

We now turn our attention to the invertibility aspect, initiating our discussion with the special case where $d = 1$.

In this one-dimensional setting, the probability that the matrix $\frac{1}{n_{\mathrm{I_{opt}}}} \sum_{i=1}^{n_{\mathrm{I_{opt}}}} X_i X_i^T$ fails to be invertible (i.e., is singular) can be reformulated as follows:

$$P\left\{ \frac{1}{n_{\mathrm{I_{opt}}}} \sum_{i=1}^{n_{\mathrm{I_{opt}}}} X_i X_i^T \right\} = P\left\{ \frac{1}{n_{\mathrm{I_{opt}}}} \sum_{i=1}^{n_{\mathrm{I_{opt}}}} (X_i)^2 = 0 \right\}$$

Here, we have implicitly assumed that X_i are scalar random variables (as $d = 1$) and thus $X_i X_i^T$ reduces to $(X_i)^2$.

Given that the sum of non-negative terms can only be zero if each individual term is zero, we have:

$$P\left\{ \tfrac{1}{n_{\mathrm{I_{opt}}}} \textstyle\sum_{i=1}^{n_{\mathrm{I_{opt}}}} (X_i)^2 = 0 \right\}$$
$$= P\left\{ (X_i)^2 = 0, \forall i \in \{1, \ldots, n_{\mathrm{I_{opt}}}\} \right\}$$

This probability is further bounded above by the probability that any single term is zero, which, under the assumption of non-degenerate Gaussian distributions for X_i, is zero:

$$P\left\{ (X_i)^2 = 0, \forall i \right\} \leq P\{(X_1)^2 = 0\} = 0$$

Hence, the probability that the matrix $\frac{1}{n_{\mathrm{I_{opt}}}} \sum_{i=1}^{n_{\mathrm{I_{opt}}}} X_i X_i^T$ is singular in the case $d = 1$ is zero, as desired.

This concludes our proof of the invertibility and symmetry properties under the specified conditions.

Consider now the scenario where $d > 1$. Take the first k outcomes, denoted as $X^1, \ldots, X^{\mathrm{I_{opt}}}$. These vectors span an affine subspace of dimension at most k within $\mathbb{R}^{n_{\mathrm{I_{opt}}}}$. Importantly, this affine subspace constitutes a null set under any non-degenerate Gaussian distribution on $\mathbb{R}^{n_{\mathrm{I_{opt}}}}$.

Next, we examine the probability of a specific event related to these outcomes. Specifically, we consider the probability that the sample covariance matrix, $\frac{1}{n_{\mathrm{I_{opt}}}} \sum_{i=1}^{n_{\mathrm{I_{opt}}}} X_i X_i^T$, is singular. This can be formulated as follows:

$$P\left\{ \tfrac{1}{n_{\mathrm{I_{opt}}}} \textstyle\sum_{i=1}^{n_{\mathrm{I_{opt}}}} X_i X_i^T \right\} =$$

$$P\left\{ \exists a \in \mathbb{R}^{n_{\mathrm{I_{opt}}}} \setminus \{0_{n_{\mathrm{I_{opt}}}}\} \colon a^T \left(\tfrac{1}{n_{\mathrm{I_{opt}}}} \textstyle\sum_{i=1}^{n_{\mathrm{I_{opt}}}} X_i X_i^T \right) a = 0 \right\}$$

This simplifies to:

$$= P\left\{ \min_{a \in \mathbb{R}^{n_{\mathrm{I_{opt}}}} \setminus \{0_{n_{\mathrm{I_{opt}}}}\}} \frac{1}{n_{\mathrm{I_{opt}}}} \sum_{i=1}^{n_{\mathrm{I_{opt}}}} (a^T X_i)^2 = 0 \right\}$$

which is equivalent to saying that there exists a non-zero vector a such that a is orthogonal to all X_i for $i = 1, \ldots, n_{\mathrm{I_{opt}}}$:

$$= P\left\{ \exists a \in \mathbb{R}^{n_{\mathrm{I_{opt}}}} \setminus \{0_{n_{\mathrm{I_{opt}}}}\} \colon a \perp X_i \right\}$$

for all $i \in \{1, \ldots, n_{\mathrm{I_{opt}}}\}$.

Since the dimension of the affine subspace spanned by $X^1, \ldots, X^{\mathrm{I_{opt}}}$ is at most k (which is at most $d - 1$ in our context, assuming $k \leq d - 1$), the probability of finding such a non-zero orthogonal vector a is further bounded by considering the case where $k = d - 1$:

$$\leq P\left\{ \exists a \in \mathbb{R}^{n_{\mathrm{I_{opt}}}} \setminus \{0_{n_{\mathrm{I_{opt}}}}\} \colon a \perp X_i \right\}$$

for all $i \in \{1, \ldots, d\}$.

However, due to the properties of Gaussian distributions and the fact that the Lebesgue measure dominates all non-degenerate Gaussian distributions, we conclude that the event described above—a non-zero vector a being orthogonal to all X_i—has probability zero. Hence:

$$P\left\{ \frac{1}{n_{\mathrm{I_{opt}}}} \sum_{i=1}^{n_{\mathrm{I_{opt}}}} X_i X_i^T \right\} = 0$$

This proves that affine subspaces of dimension at most $d-1$ in $\mathbb{R}^{n_{\mathrm{I_{opt}}}}$ are null sets under any non-degenerate Gaussian distribution. □

III. SIMULATION

A. Preparatory work

1) Index: To evaluate the accuracy of predictions in data simulation, the MSE and MAE are utilized to measure the difference between the true values and the estimated values. The formulas for MSE and MAE, which quantify the errors, are given as follows:

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

2) Redundant data preparation: We build the (X, Y) is from the model $Y_i = X_i\beta + \varepsilon_i$, we can know X consists of (X_1, X_2) and Y consists of (Y_1, Y_2). we can define as:

$$X_1 = (X_{1ij}) \in \mathbb{R}^{n_1 \times p}, X_{1ij} \sim N(0, 2).$$

$$Y_1 = X_1\beta + \varepsilon_1, n_1 \in (1, ..., n - n_r).$$

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

$$Y_2 = X_2\beta + \varepsilon_2, n_2 \in (1, ..., n_r).$$

we can know $\beta \sim Unif(0, 2)$ and $\varepsilon = (\varepsilon_1, \varepsilon_2)$, where $\varepsilon_1 \sim N(0, 3), \varepsilon_2 \sim N(0, 5)$, and run our simulation.

B. Simulation analysis

The steps of the simulation analysis include stability analysis and sensitivity analysis. Stability analysis involves observing the variations of n and p, while sensitivity analysis focuses on the variations of K and n_r. Then, selected simulation data from three different distributions to conduct experiments, resulting in the following three cases.

Case 1. $X_2 = \left(X_{ij}\right) \in \mathbb{R}^{n_2 \times p},\ X_{2ij} \sim Po(6)$.

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

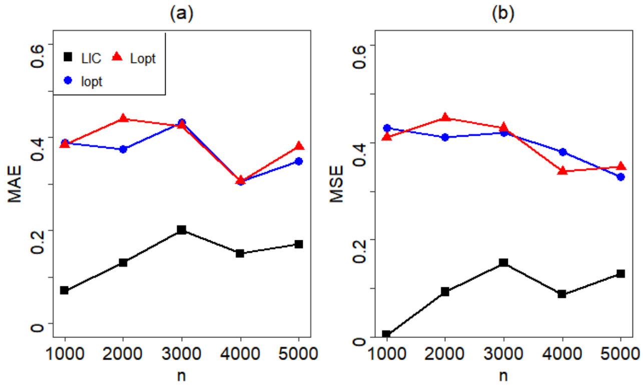

Fig. 1. Vary the value of n leads to changes in MAE and MSE.

From Figure 1, with fixed K, p, n_r, and vary n, observe that the trends of MAE and MSE are similar. Firstly, it is evident that among LIC, Lopt and Iopt, LIC has the lowest MAE and MSE, indicating the superior performance of the LIC method. Subsequently, as n varies, we see an initial increase in the curves. When n increases from 3000 to 4000, MAE decreases from 0.212 to 0.156, while MSE decreases from 0.152 to 0.088. However, as n further increases from 4000 to 5000, MAE increases from 0.156 to 0.171, and MSE rises from 0.088 to 0.130. Therefore, we determine that the optimal simulation state is achieved when n is set to 4000.

Scenario 2: Setting $K = 10$, $n = 3000$, $\alpha = 0.05$, $n_r = 50$, vary $p = 8, 9, 10, 11, 12$.

From Figure 2, with fixed K, n, n_r, and vary p, observe that the trends of the MAE and MSE curves generated by fitting are similar. However, the trend of the LIC curve is decreasing. As p reaches 10, we can see that the values of MAE and MSE reach their minima, which are 0.072 and 0.059, respectively. As p increases to 11, we find that MAE and MSE change to 0.12 and 0.079, respectively. Therefore, the optimal simulation state is achieved as p is set to 10.

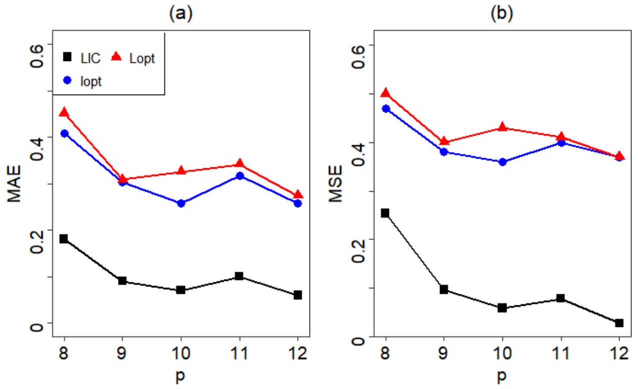

Fig. 2. Vary the value of p leads to changes in MAE and MSE.

Combining Scenario 1 and Scenario 2, the comparison reveals that the MAE and MSE exhibit significantly similar trends, both being smaller compared to the other two algorithms (Iopt and Lopt). This demonstrates the good stability of the LIC algorithm. Detailed numerical analyses for Scenario 1 and Scenario 2 are provided below.

Scenario 3: Setting $n = 6000$, $p = 8$, $\alpha = 0.05$, $n_r = 50$, vary $K = 5, 10, 15, 20, 25$.

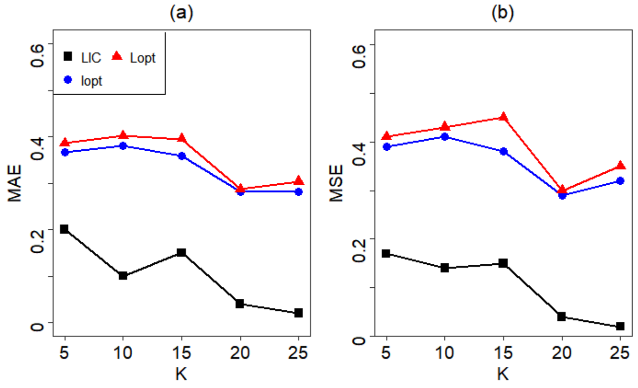

Fig. 3. Vary the value of K leads to changes in MAE and MSE.

From Figure 3, with fixed n, p, n_r and vary K, observe distinct differences in the trends of the MAE and MSE fitting curves. Firstly, considering the MAE trend, it exhibits a downward trend as K increases from 5 to 10, but then an upward trend from 10 to 15. The specific MAE values are 0.103, 0.174 and 0.046, respectively. As for the MSE trend, the MSE curve generally displays a downward trend. Notably, when K is 10, we identify it as an inflection point of the downward trend, with an MSE value of 0.171. Therefore, the optimal simulation state is achieved as K is set to 10.

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

From Figure 4, with fixed n, p, K, and vary n_r, observe that the trends of the MAE and MSE fitting curves are similar. As n_r changes from 40 to 50 and then to 60, the degree of change is minimal. Specifically, the MAE values are 0.072, 0.061 and 0.057, respectively, while the MSE values are 0.104, 0.099 and 0.095, respectively. However, as n_r increases from 60 to 70, the MAE and MSE values change significantly, to 0.057 and 0.099, respectively. Therefore, we conclude that the optimal simulation state is achieved as n_r

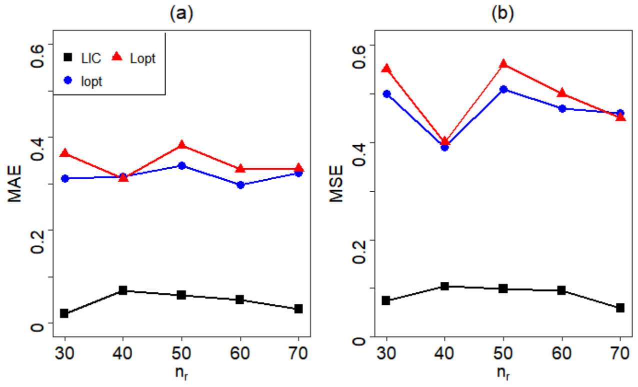

Fig. 4. Vary the value of n_r leads to changes in MAE and MSE.

is 60.

Combining Scenario 3 and Scenario 4, the values of MAE and MSE for the LIC algorithm are relatively large, indicating poor model performance. As K increases, an inflection point emerges when $K = 15$, where the values of MAE and MSE begin to decrease, and the performance starts to improve. Observe the value of n_r, performance starts to enhance after n_r reaches 60. Detailed numerical analyses are presented in Scenario 3 and Scenario 4.

In summary, we have obtained the optimal values for each variable in fitting the Po distribution to the data. When $\{n, p, K, n_r\}$ are set to $\{4000, 10, 10, 60\}$, the LIC achieves the best fitting state. Therefore, under these conditions, we can obtain the optimal subset.

We simulate a data set following a Poisson distribution with $\lambda = 6$. This means that, on average, six events occur in each time interval.

In the simulation process, we employ the LIC criterion to assess the redundancy of the data. The LIC criterion is a statistical method designed to determine the presence of redundant variables within a dataset. It rests on the premise that when two or more variables exhibit a high degree of correlation, they are likely to be redundant.

In our simulation, we calculate the LIC value for the dataset and utilized this metric to identify the presence of redundant variables. We discovered that, in the Poisson distribution with $\lambda = 6$, all observations are independent, indicating the absence of redundancy. Consequently, in this particular simulation scenario, there are no redundant variables.

Our simulation results indicate that for the Poisson distribution with $\lambda = 6$, the variables in the dataset are not redundant. This is likely due to the fact that in this specific Poisson distribution, each event is independent and there is no dependency between events. However, this does not imply that redundancy is absent in all Poisson distributions. In other Poisson distributions, if the average incidence rate of events is relatively high or low, or if there exists some form of dependency between events, redundant variables may emerge. Therefore, it is crucial to utilize the LIC criterion to evaluate redundant variables in datasets in practical applications. Through this simulation study, we have gained a deeper understanding of how to apply the LIC criterion to assess redundant variables in Poisson distributions. This aids us in more precisely identifying and handling redundant variables in actual data analysis, thereby enhancing the accuracy and reliability of the analysis.

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

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

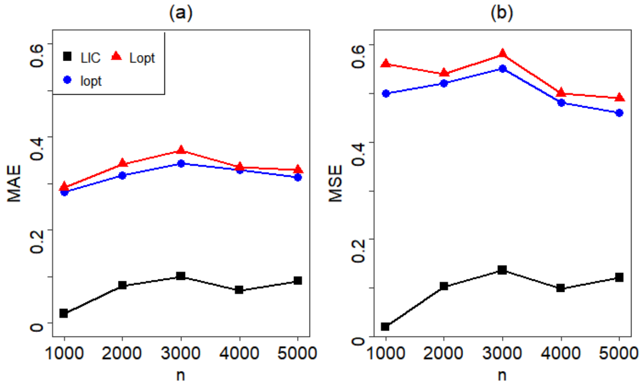

Fig. 5. Vary the value of n leads to changes in MAE and MSE.

From Figure 5, with fixed K, p and n_r and vary n, observe that the trends of MAE and MSE are similar. As the value of n increases, both MAE and MSE gradually increase. Specifically, as n increases from 3000 to 4000, we find that MAE decreases from 0.108 to 0.075, while MSE also decreases from 0.137 to 0.099. However, as n increases from 4000 to 5000, both MAE and MSE begin to gradually increase again. Therefore, we conclude that the optimal simulation state is achieved as n is set to 4000.

Scenario 2: Setting $K = 10$, $n = 3000$, $\alpha = 0.05$, $n_r = 50$, vary $p = 8, 9, 10, 11, 12$.

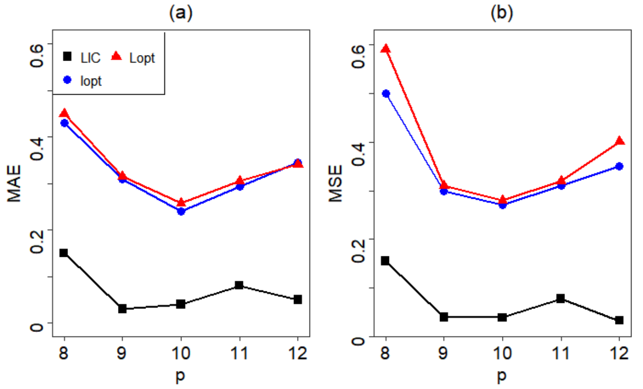

Fig. 6. Vary the value of p leads to changes in MAE and MSE.

From Fig. 6, with fixed K, n, n_r and vary p, observe that the trends of MAE and MSE are similar. As the value of p increases, the values of MAE and MSE gradually decrease. As p reaches 9, both MAE and MSE reach their minimum values, which are 0.033 and 0.039 respectively. However, as the value of p continues to increase, the values of MAE and MSE gradually increase. Therefore, the optimal simulation state is achieved as p is equal to 10.

Combining Scenario 1 and Scenario 2, a comparison of the MAE and MSE values among the three algorithms reveals that they exhibit the same trend as n increases. Moreover, the LIC algorithm yields the smallest MAE and MSE values, further confirming its excellent stability. Detailed numerical analyses are presented in Scenario 1 and Scenario 2.

Scenario 3: Setting $n = 3000$, $p = 8$, $\alpha = 0.05$, $n_r = 50$, vary $K = 5, 10, 15, 20, 25$.

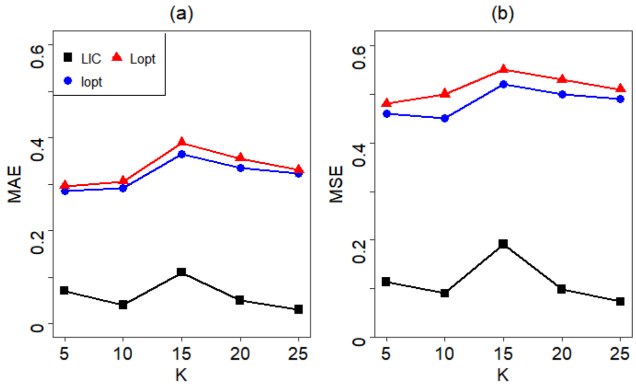

Fig. 7. Vary the value of K leads to changes in MAE and MSE.

From Fig. 7, with fixed n, p, n_r and vary K, observe that the changing trends of MAE and MSE are similar. As K varies from 5 to 10, MAE and MSE reach their minimum values. However, as K increases from 10 to 15, MAE and MSE begin to increase. As they reach their minimum values, we can see that MAE and MSE are 0.045 and 0.091, respectively. Therefore, the optimal simulation state is achieved as K is 10.

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

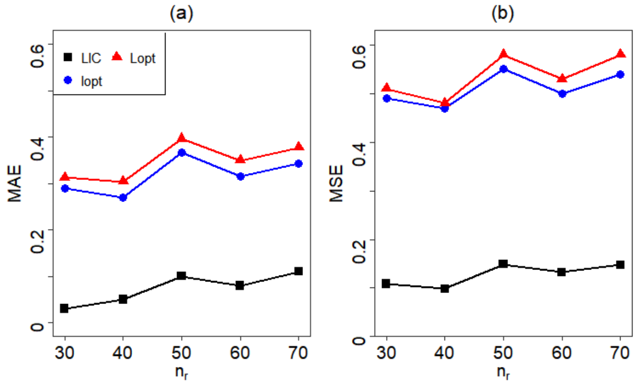

Fig. 8. Vary the value of n_r leads to changes in MAE and MSE.

From Fig. 8, with fixed n, p, K and vary n_r, observe that the changing trends of MAE and MSE are similar. As n_r gradually increases, MAE also gradually increases, decreases slightly from 50 to 60, and gradually increases again from 60 to 70. Therefore, the optimal value of MAE is 0.088 as n_r is 60. MSE first decreases slightly and then increases, gradually decreasing from 50 to 60 and gradually increasing from 60 to 70. As a result, the optimal value of MSE is 0.132 as n_r is 60. Therefore, the best simulation state is achieved as n_r is 60.

Combining Scenario 3 and Scenario 4, as the number of partitions K increases, the performance of the LIC algorithm initially deteriorates, reaching maximum MAE and MSE values when K reaches 15. Subsequently, the performance gradually improves. Comparative analysis reveals that the best performance is achieved when $K = 10$. Meanwhile, observing the nr value, optimal performance is also attained when $n_r = 60$. In general, both the number of partitions K and the variation of the parameter n_r significantly impact the overall performance of the model. Detailed numerical analyses are provided in Scenario 3 and Scenario 4.

In summary, we have obtained the optimal values of various variables for fitting data with the exponential distribution. When $\{n, p, K, n_r\}$ are set as $\{4000, 10, 10, 60\}$, the LIC reaches its best fitting state. Therefore, we can obtain the optimal subset under these conditions.

To simulate this distribution, we first need to generate a set of data. This data set will be based on the exponential distribution with $\theta = 3$. Using this method, we can simulate a random data set that conforms to specific parameter settings.

In the simulation process, we employ the LIC criterion to analyze the data. The LIC criterion is a method for evaluating data redundancy, assisting us in determining whether there is excessive duplicate information in the data set. By applying the LIC criterion, we can gain insight into the structure and patterns of the data set and decide whether data dimension reduction or simplification processing is required.

The simulation results will indicate whether the data set contains excessive redundant information. If the redundancy is high, further processing of the data may be necessary to reduce duplicate information. If the redundancy is low, the data set may already be a relatively concise and effective representation. Additionally, by comparing simulation results under different parameter settings, we can gain a deeper understanding of how parameters affect data redundancy.

In summary, by simulating data with an exponential distribution and analyzing it using the LIC criterion, we can better understand the level of data redundancy and take appropriate measures accordingly. This is crucial during the data analysis and preprocessing stages, as it helps us identify and address unnecessary data redundancy, thus enhancing the efficiency and accuracy of the analysis.

Case 3. $X_2 = \left(X_{ij}\right) \in \mathbb{R}^{n_2 \times p},\ X_{2ij} \sim NBin(10, 1)$.

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

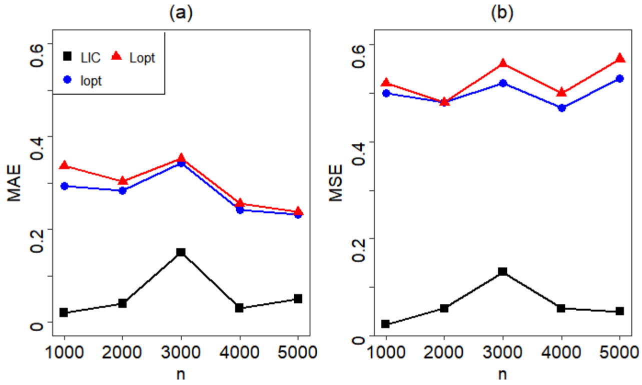

Fig. 9. Vary the value of n leads to changes in MAE and MSE.

From Fig.9, with fixed K, p and n_r and vary n, observe that the trends of MAE and MSE are similar. As the value of n increases, the values of MAE and MSE also gradually increase. As n changes from 3000 to 4000, MAE decreases from 0.153 to 0.034, and MSE decreases from 0.131 to 0.056, reaching their minima. As the value of n continues to increase, MAE and MSE gradually increase. Therefore, we determine that the optimal simulation state is achieved when n is set to 4000.

Scenario 2: Setting $K = 10$, $n = 3000$, $\alpha = 0.05$, $n_r = 50$, vary $p = 8, 9, 10, 11, 12$.

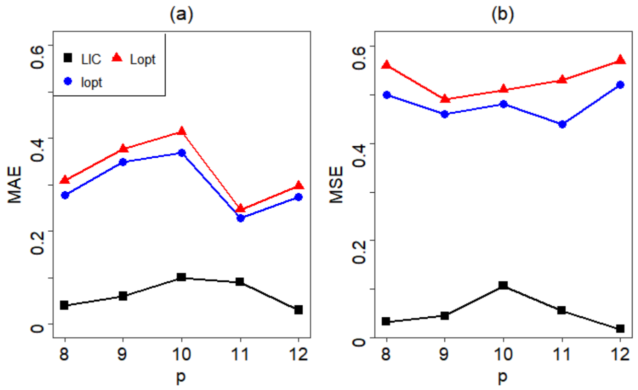

Fig. 10. Vary the value of p leads to changes in MAE and MSE.

From Fig. 10, with fixed K, n and n_r and vary p, observe that the trends of MAE and MSE are similar. As the value of p increases, both MAE and MSE initially increase and then decrease. The minimum values of MAE and MSE are achieved when p is 12, with values of 0.031 and 0.018, respectively. Therefore, the optimal simulation state is achieved as p is set to 12.

Combining Scenario 1 and Scenario 2, the trends observed for the three algorithms show slight variations but overall consistency. Notably, the LIC algorithm consistently maintains the lowest MAE and MSE values compared to the other two algorithms, which further demonstrates its excellent stability. Detailed numerical analyses for both Scenario 1 and Scenario 2 are provided.

Scenario 3: Setting $n = 3000$, $p = 8$, $\alpha = 0.05$, $n_r = 50$, vary $K = 5, 10, 15, 20, 25$.

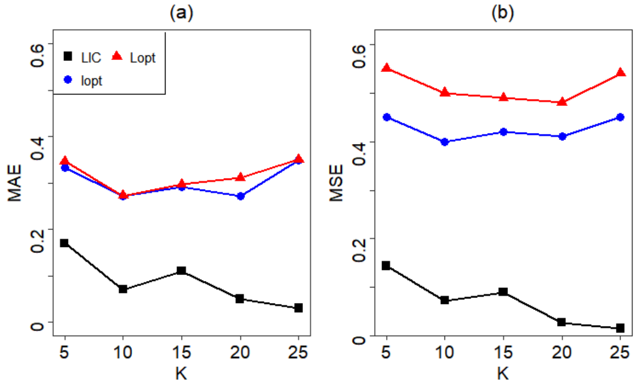

Fig. 11. Vary the value of K leads to changes in MAE and MSE.

From Fig. 11, with fixed n, p, n_r and vary K, observe that the change trends of MAE and MSE are similar. As the value of K ranges from 5 to 10, MAE and MSE gradually decrease. As the value of K ranges from 10 to 15, MAE and MSE gradually increase. As the value of K is greater than 15, MAE and MSE show a decreasing trend. As the value of K is equal to 10, MAE and MSE are 0.071 and 0.073, respectively. Therefore, the best simulation state is achieved when the value of K is 10.

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

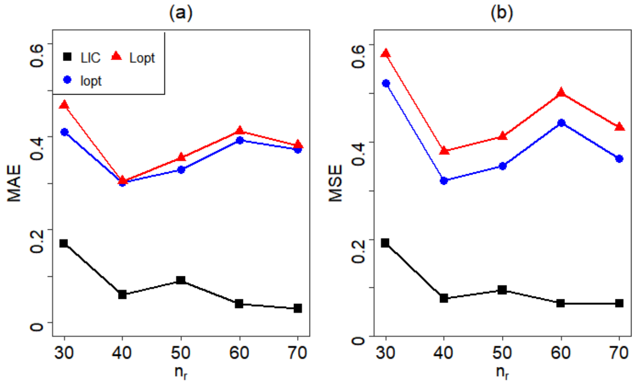

Fig. 12. Vary the value of n_r leads to changes in MAE and MSE.

From Fig. 12, with fixed n, p, K and vary n_r, observe that the change trends of MAE and MSE are similar. We can see that from 30 to 40, MAE and MSE gradually decrease, and from 40 to 50, MAE and MSE gradually increase. We believe that n_r achieves the best simulation effect at 40, with MAE and MSE being 0.063 and 0.078, respectively.

Combining Scenario 3 and Scenario 4, the analysis of K and n_r reveals that the performance of the LIC algorithm is significantly superior to the other two algorithms. For the negative binomial distribution, the LIC algorithm achieves optimal model performance when $K = 10$ and $n_r = 40$. This further underscores that both the number of partitions K and the variation of parameter n_r significantly influence the overall performance of the model. Detailed numerical analyses for both Scenario 3 and Scenario 4 are provided.

In summary, we obtain the optimal values of various variables for fitting the data with the negative binomial distribution. When $\{n, p, K, n_r\}$ are respectively $\{4000, 12, 10, 60\}$, the LIC achieves the best fitting state. Therefore, we can obtain the optimal subset under this condition.

In the simulation process, we generate a random data set with a negative binomial distribution. The size of the data set is n, and each observation x originates from a latent random variable Y, whose distribution parameters are r and p. In the simulation, we simulate different shapes of the negative binomial distribution by adjusting the values of r and p.

During the simulation, we apply the LIC criterion to analyze the generated simulation data. Initially, we calculate the similarity or correlation between each observation x and the other observations in the data set based on the available data. Subsequently, we utilize this similarity or correlation information to assess the redundancy level within the data set. Specifically, we evaluate the redundancy between different data blocks by comparing their similarity or correlation. If two data blocks exhibit high similarity or correlation, it is considered that there is redundant information between them.

Through simulation analysis, we reach some interesting conclusions. Firstly, as the observations in the dataset come from a negative binomial distribution with a larger r value, there is less redundant information in the dataset. This is probably because a larger r value indicates that each observation is more independent, without too much duplicate information. Conversely, as r is small, there is more redundant information in the dataset, as there is more duplicate information between the observations. In addition, we also found that when p is close to 1, there is less redundant information in the dataset. This is because when p is close to 1, each observation is more likely to succeed rather than fail, thus reducing the occurrence of duplicate information.

By applying the LIC criterion, we can gain a better understanding of the redundant information in negative binomial distribution datasets. This helps us better handle redundant information in practical data analysis, such as improving data usability and reliability through data dimension reduction or simplified processing. Furthermore, this method can be extended to other types of distributions and models, further enriching our understanding and processing capabilities for different types of data.

IV. CONCLUSION

In a distributed data environment, different data sources may contain duplicate or highly correlated information, which increases redundancy. Selecting the optimal subset can reduce redundancy and improve the efficiency of estimation. When selecting the optimal subset, the size of the subset must be considered. A smaller subset may reduce redundancy, but it may also lead to information loss or instability in estimation. A larger subset may provide more information, but redundancy will also increase. Therefore, there is an optimal subset size that achieves a balance between estimation redundancy and information content. Formulating clear criteria for selecting the optimal subset is necessary. These criteria should be determined by the fundamental properties of the data, the goals of the analysis, and the accuracy requirements for the estimation. Common criteria include AIC, BIC, cross-validation, etc., which can help evaluate the predictive and explanatory abilities of different subsets.

When determining the optimal subset, it is necessary to choose an appropriate statistical model. The choice of model should be based on the distribution of the data, the dimensionality of the data, and the purpose of the analysis. For example, linear regression, logistic regression, decision trees, random forests, etc., may all be applicable in different scenarios. In a distributed data environment, there may be interactions between different data sources that affect the selection of the optimal subset. Considering interactions can help to more fully understand the relationships between data and may reveal hidden patterns.As time passes, data may change, requiring dynamic updating of the optimal subset. Continuously monitoring changes in data and adjusting subsets can ensure the timeliness and accuracy of estimation. When selecting the optimal subset, consideration should be given to the interpretability of the results. Complex models or subsets that are difficult to interpret may limit their application in decision-making. Balancing complexity and interpretability is crucial.

Ensuring that the selected optimal subset has good generalizability is essential. Generalizability refers to a model's ability to perform well on new data. Appropriate validation and cross-validation techniques can be used to assess and improve a model's generalizability. Proper data preprocessing is crucial before selecting the optimal subset. This includes cleaning up duplicates, missing or outlier values, and possible feature scaling or normalization. These steps can contribute to enhancing the accuracy and stability of the estimation. For the selected subset, appropriate evaluation metrics should be used for validation. These metrics should be relevant to the analyzed problem and provide useful feedback on model performance.

In summary, estimating an optimal subset for redundant distributed data is a multifaceted problem that requires consideration of the nature of the data, the purpose of analysis, and statistical principles. By delving into these issues, more accurate conclusions can be drawn and applied in practical data analysis.

V. OUR FURTHER WORK

In practical applications, to enhance the realism and accuracy of simulation results, we can flexibly set block length and block number in an unconstrained manner, enabling the data to more accurately simulate actual situations. Simultaneously, based on our needs, we can replace redundant distributions with alternative distributions, such as high-dimensional models or probabilistic models, to better adapt to different scenarios and requirements. This flexible simulation approach can provide more meaningful references and guidance for distributed data processing, assisting us in better understanding and processing actual data.

In future work, we will further investigate the potential of unconstrained settings for block length and block number in simulations, as well as the substitution of redundant distributions with alternative ones. These adjustments will enhance the flexibility of our method in handling distributed data. Additionally, we will consider replacing existing models with other types, including high-dimensional models and probabilistic models, to broaden our application scenarios and achieve more accurate results. Through these improvements, we anticipate making more significant contributions to the field of distributed data processing.

DATA AVAILABILITY

We have employed the LIC criterion to fit the relevant data matrices of Poisson distribution, exponential distribution, and negative binomial distribution respectively, thus simulating distributed redundant data to study the optimization of the LIC criterion for distributed redundant data. We have packaged the implemented LIC criterion into an R package called LIC, which is available for download at the following URL: https://CRAN.R-project.org/package=LIC.

REFERENCES

[1] 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.

[2] G. Guo, "Parallel statistical computing for statistical inference," *Journal of Statistical Theory and Practice*, vol. 6, pp. 536–565, 2012.

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

[5] 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.

[6] J. Lederer. “Fundamentals of High-Dimension Statistics,” *Switzerland. Springer Nature Switzerland*, AG. 2020. 1.

[7] 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.

[8] G. Guo, Y. Sun, G. Qian, and Q. Wang. “LIC: The LIC Criterion for Optimal Subset Selection.” 2022.

[9] 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.

[10] G. Guo, C. Wei, and G. Q. Qian, “SOPC: The Sparse Online Principal Component Estimation Algorithm,” 2022.

[11] 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.

[12] 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.

[13] 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.

[14] 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.