An Optimal Subset Selection Algorithm for Distributed Hypothesis Test

Jiarui Li, Guangbao Guo

***Abstract*—Big data has made analyzing redundant and distributed data a significant challenge. This article presents a new method to determine the optimal subset of redundant distributed data. This method is based on PPC algorithms, which allow it to extract valuable insights from redundant data, making it easier to estimate the optimal data subset. Testing has shown that this method improves data quality and enhances data utilization and performance evaluation.**

***Index Terms*—Distributed redundant data, optimal subset test, PPC method, performance evaluation.**

I. Introduction

THE novelty of this research is the selection of an optimal subset. Our method differs from past methods that may include extra information. We select the subset using the intersection of two estimation algorithms, which forms our optimization scheme. An advantage of this method is that it uses a small amount of data to contain a significant amount of information, leading to much better outcomes than previous work.

Research in distributed data explores divide-and-conquer algorithms. It studies communication strategies to optimize big data statistics. Two communication algorithms are examined. Optimal subset selection is a strategy for managing big data. A method called leverage is proposed for subsampling.

The research explores principal component analysis and algorithmic aspects for data distributed across servers. Variational Bayesian formulations are applied to derive optimal probability distributions for regression parameters. Other advancements include partitioned quasi-likelihood, parallel statistical computing, parallel maximum likelihood estimation, and distributed online EM. These contribute to efficient and accurate insights from big data. The article proposes a distributed algorithm PPC for subset selection and prediction in linear models. It uses LIC as a comparison method.

In the realm of distributed statistical inference, Guo et al.[1] introduced LIC method, which has enhanced the efficiency of statistical inference in big data environments. Guo et al. [2] devised an optimization program for distributed interval estimation problems, leading to improved estimation accuracy. The literature also includes references to Guo et al. [3]–[5] and Ma et al. [6]. These methodologies collectively provide effective statistical inference tools for handling large-scale distributed data.

Manuscript received March 10, 2024; revised November 20, 2024.

This project received funding from the National Social Science Foundation Project (ID: 23BTJ059), the Natural Science Foundation of Shandong (ID: ZR2020MA022), and the National Statistical Research Program (ID: 2022LY016).

Jiarui Li is a postgraduate student of Mathematics and Statistics, Shandong University of Technology, Zibo, China. (e-mail: jiarui0191 @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).

A. Linear model

The model utilizing multiple linear regression is

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_K X_K + \varepsilon.$$

In practice, we employ a distributed method to divide the original data matrix into K blocks by utilizing a sequence of projection matrices $\{R_k\}_{k=1}^{K}$. The process of partitioning the original data matrix into K blocks is achieved through the following transformations

$$Y_{I_k} = R_k Y, X_{I_k} = R_k X, \varepsilon_{I_k} = R_k \varepsilon,$$

where Y_{I_k} and X_{I_k} represent the $n_{I_k} \times p$ submatrices of Y and X, $n_{I_k} \geq p$. Additionally, ε_{I_k} denotes the sub-residual vector. The original data is divided into K blocks using distributed computing methods. Since this process is conducted independently on the K machines, each machine yields a processed data subset. Therefore, after applying the optimization method, the data subsets residing on the K machines can be expressed as $Q_{I_k} = (Y_{I_k}, X_{I_k})$, Q_{I_k} denotes the data subset on the $k-th$ machine, where Y_{I_k} and X_{I_k} represent the response variable and feature variables of that subset, respectively. The subscript I_k indicates that this is the $k-th$ subset and implicitly suggests that each subset may contain only a portion or specific part of the original dataset Q. In summary, the entire dataset $Q = (Y, X) = \{(Y_{I_k}, X_{I_k})\}_{k=1}^{K}$ is segmented into K parts and processed in parallel on K machines.

B. Our work

Firstly, the PPC algorithm is introduced as a means to filter the best subsets by intersecting the subsets derived from two specified criteria. This algorithm boasts the advantage of shortening the subset length without compromising estimation accuracy, thereby enhancing overall work efficiency.

Secondly, the study examines how explanatory variables, sample size, and dimensionality affect the PPC. Results show that the PPC method's estimation accuracy improves with larger sample sizes. Additionally, as dimensionality increases, PPC's estimation accuracy also rises. A comparison between PPC and LIC methods further confirms PPC's effectiveness and superiority.

Ultimately, the proposed method not only enhances dimensionality reduction efficiency but also boosts estimation accuracy while significantly decreasing computational load.

We have developed the code into an R package named **PPCDT**.

II. Method and Theorem

The model is $Y_{I_k} = X_{I_k}\beta + \varepsilon_{I_k}$ where Y_{I_k} is the random response variable with observations distributed as

$Y_{I_k} \sim N\left(\mu_{I_k}, \sigma_{I_k}^2\right), k = 1, ..., K$, where $\hat{\mu}_{I_k} = X_{I_k}\hat{\beta}_{I_k}$ and $\hat{\beta}_{I_k} = \left(X_{I_k}^T X_{I_k}\right)^{-1} X_{I_k}^T Y_{I_k}$. The estimator is

$$\hat{\sigma}_{I_k}^2 = \frac{1}{n_{I_k} - P_n}\left(Y_{I_k} - X_{I_k}\hat{\beta}_{I_k}\right)^T\left(Y_{I_k} - X\hat{\beta}_{I_k}\right)$$

$$= \frac{1}{n_{I_k} - P_n} Y_{I_k}^T (I_{I_k} - H_{I_k}) Y_{I_k}.$$

For the projection matrix $\{R_k\}_{k=1}^K$, it is obtained that $Y_{I_k} = R_k Y, X_{I_k} = R_k X, \varepsilon_{I_k} = R_k \varepsilon$. The distributed hypothesis testing is

$$H_{I_k 0}{:}A_{I_k}\beta \leq b_{I_k}, H_{I_k 1}{:}A_{I_k}\beta > b_{I_k}$$

where $A_{I_k} = R_k A$, $b_{I_k} = R_k b$, then

$$\hat{\beta}_{I_k}(\hat{H}_{I_k 0}) = \hat{\beta}_{I_k} -$$

$$(X_{I_k}^T X_{I_k})^{-1} A_{I_k}^T \left(A_{I_k}(X_{I_k}^T X_{I_k})^{-1} A_{I_k}^T\right)^{-1}\left(A_{I_k}\hat{\beta}_{I_k} - b_{I_k}\right),$$

$$\hat{\sigma}_{I_k}^2\left(H_{I_k 0}\right) =$$

$$\frac{1}{n_I - P_n}\left(Y_{I_k} - X_{I_k}\hat{\beta}_{I_k}\left(H_{I_k 0}\right)\right)^T\left(Y_{I_k} - X_{I_k}\hat{\beta}_{I_k}(H_{I_k 0})\right).$$

In the case where the null hypothesis holds true, the t-test statistic is

$$T_{I_k} = \frac{(A_{I_k}\hat{\beta}_{I_k} - b_{I_k})/(\sigma_{I_k}/\sqrt{n_{I_k}})}{\sqrt{S_{I_k}^2/\sigma_{I_k}^2}}.$$

The F-test statistic would be expressed as

$$F_{I_k} = \frac{(\widehat{\sigma}_{I_k}^2(H_{I_k 0}) - \widehat{\sigma}_{I_k}^2)/(\sigma_{I_k}^2 \cdot m)}{\widehat{\sigma}_{I_k}^2/(\sigma_{I_k}^2 \cdot (n_{I_k} - P_n))} \sim F_{I_k}(m, n_{I_k} - P_n)$$

where $S_{I_k}^2$ is the variance of the $k-th$ block of the sample matrix.

Different tests can lead to different conclusions for a given sample. Test power is a crucial metric for assessing the effectiveness of a test, and a higher value indicates greater efficiency. Assuming the rejection region of the test is W, and the sample observations are X, the power of the test represents the probability of correctly rejecting H_0 when H_1 is true. The power $g(\beta)$ can be written as $1 - \gamma(\beta)$ where $\gamma(\beta)$ is the probability of accepting H_0 when β belongs to H_1:

$$g(\beta) = 1 - \gamma(\beta) = 1 - P_\beta(accept H_0 | \beta \in H_1).$$

For the hypothesis

$$H_{I_k 0, \sigma_{I_k}^2} : \sigma_{I_k}^2(H_{I_k 0}) \leq \sigma_{I_k}^2, H_{I_k 1, \sigma_{I_k}^2} : \sigma_{I_k}^2(H_{I_k 0}) > \sigma_{I_k}^2,$$

where $\sigma_{I_k}^2(H_{I_k 0}) = \sigma_a^2 > \sigma_{I_k}^2$, then

$$F_{I_k} = \frac{(\widehat{\sigma}_{I_k}^2\left(H_{I_k 0}\right) - \widehat{\sigma}_{I_k}^2)/(\sigma_{I_k}^2 \cdot m)}{\widehat{\sigma}_{I_k}^2/(\sigma_{I_k}^2 \cdot (n_{I_k} - P_n))} \sim F_{I_k}(m, n_{I_k} - P_n).$$

The probability is

$$\gamma\left(\sigma_{I_k}^2(H_{I_k 0})\right) = P(F_{I_k} < F_\alpha(m, n_{I_k} - P_n) | \sigma_a^2 > \sigma_{I_k}^2).$$

Therefore, the power of the test is

$$g\left(\sigma_{I_k}^2(H_{I_k 0})\right) = 1 - \gamma\left(\sigma_{I_k}^2(H_{I_k 0})\right) = 1 - P(0 < F_{I_k} < \frac{\sigma_{I_k}^2}{\sigma_a^2} F_\alpha).$$

For the subset sequence $\{I_k\}_{k=1}^K$, the optimal indicator subset based on maximum power is

$$I_F^1 = \arg\max_{I_k}\left\{1 - P(0 < F_{I_k} < \frac{\sigma_{I_k}^2}{\sigma_a^2} F_\alpha)\right\}.$$

The hypothesis may be useful to consider that for a particular parameter value μ_a. The t-test statistic is

$$T_{I_k} = \frac{\bar{X}_{I_k} - \mu_a}{S_{I_k}/\sqrt{n_{I_k}}}, T_{I_k} \sim N(0, 1).$$

Similarly, for the subset sequence $\{I_k\}_{k=1}^{K_n}$, we select the optimal indicator subset based on maximum power

$$I_t^1 = \arg\max_{I_k}\left\{1 - \phi\left(t_a + \frac{\mu_{I_k 0} - \mu_a}{s_{I_k}/\sqrt{n_{I_k}}}\right)\right\}.$$

When presenting the outcomes of hypothesis testing, it is not only beneficial to rely on power but also to incorporate the p-value. Furthermore, for each sample point (x, y), a statistic $W(X, Y)$ is defined, and associated with this statistic is a probability $p(x, y)$, which represents the supremum of the probability P_β that $W(X, Y)$ exceeds or equals $W(x, y)$ over all possible parameter values β. $W(X, Y)$ is a statistic defined as follows for each sample point (x, y)

$$p(x, y) = \sup_\beta P_\beta(W(X, Y) \geq W(x, y)).$$

The optimal subset of indications is selected based on the minimum p-value

$$I_F^2 = \arg\min_{I_k}\left\{P\left\{\frac{(\widehat{\sigma}_{I_k}^2(H_{I_k 0}) - \widehat{\sigma}_{I_k}^2)/(\sigma_{I_k}^2 \cdot m)}{\widehat{\sigma}_{I_k}^2/(\sigma_{I_k}^2 \cdot (n_{I_k} - P_n))} > C_1\right\}\right\}.$$

The optimal subset of indicator I_t^2 based on the minimum p-value such that the p-value corresponds to the test statistic, exceeding a critical value C_2, where I_t^2 is

$$I_t^2 = \arg\min_{I_k}\left\{P\left\{\frac{\bar{X}_{I_k} - \mu_{I_k 0}}{S_{I_k}/\sqrt{n_{I_k}}} > C_2\right\}\right\}.$$

A novel method has been introduced for pinpointing optimal subsets by seamlessly merging two distinct screening criteria. This integration effectively removes any extraneous information, leading to a substantial decrease in the number of subsets. When it comes to distributed hypothesis testing, a PPC criterion based on the intersection of maximum power and minimum p-value is suggested. This criterion serves as a robust tool in determining the most suitable subset.

$$I_{FF}^{12} = I_F^1 \cap I_F^2.$$

One subset can be identified as I_t^1 and another subset that satisfies the minimum p-value as I_t^2. The optimal indication subset can be obtained by taking the intersection of these two subsets

$$I_{\mathrm{tt}}^{12} = I_t^1 \cap I_t^2.$$

In the aforementioned scenario, the intersection of two subsets, each satisfying an identical test, yielded a subset that met only one criterion. Alternatively, in the present context, it is feasible to stipulate that the two subsets independently fulfill the criteria of maximum power and minimum p-value. The subsets I_t^1 that satisfy the maximum power and I_F^2 that satisfy the test hypotheses while also adhering to the

minimum p-value can be obtained by taking the intersection of these two subsets

$$I_{tF}^{12} = I_t^1 \cap I_F^2.$$

Subsequently, the subset I_{tF}^{12} fulfills both the tests for variance and mean, as well as the maximum power and the minimum p-value. Similarly, the subset I_F^1 satisfies the test hypotheses while adhering to the minimum p-value.

The subset I_F^1 and the subset I_t^2 can be obtained by taking the intersection of these two subsets

$$I_{Ft}^{12} = I_F^1 \cap I_t^2.$$

The subsets obtained in the previous steps are as follows: Type1:I_{tF}^{12}, Type2:I_{FF}^{12}, Type3:I_{Ft}^{12}, Type4:I_{tt}^{12}.

Type3 and Type4 are theoretically superior to Type1 and Type2 having higher accuracy.

III. SIMULATION

A. Preparatory work

The dataset (X, Y) is produced by a linear model given by $Y_{I_k} = X_{I_k}\beta_{I_k} + \varepsilon_{I_k}$ where $\varepsilon_{I_k} \sim N\left(0, \sigma^2 I_{n_{I_k} \times n_{I_k}}\right)$, $k = 1, ..., K$. In the simulation process, (X_1, X_2) is used to construct X, and (Y_1, Y_2) is utilized to construct Y. The dataset has the following definition

$$Y_1 = X_1\beta_{I_k} + \varepsilon_1, \varepsilon_1 \sim N(0, 3),$$

$$Y_2 = X_2\beta_{I_k} + \varepsilon_2, \varepsilon_2 \sim N(0, 10),$$

$$X_1 = (X_{1ij}) \in R^{(n-m)\times p}, X_{1ij} \sim N(0, 5),$$

where $\beta_{I_k} \sim Unif(-3, 3)$, $\varepsilon \sim (\varepsilon_1, \varepsilon_2)$. This can be achieved by adjusting the values of $n, p, K, ratio$ to observe the changes in the simulation results, thereby determining the optimal parameters.

Scenario 1: The parameters are set as $n = 10$, $p = 8$, $\alpha = 0.05$, and $ratio = 0.05$. In this scenario, n is varied to 1000, 2000, 3000, 4000, and 5000. Subsequently, we obtain Figure 1 which illustrates the results.

Scenario 2: The parameters are set as $K = 10$, $n = 3000$, $\alpha = 0.05$, and $ratio = 0.05$, p is varied to 6, 7, 8, 9, and 10. The resulting Figure 2 shows the outcomes of these variations.

Scenario 3: The parameters are set as $n = 3000$, $p = 8$, $\alpha = 0.05$, and $ratio = 0.05$. In this scenario, K is varied to 5, 10, 15, 20, and 25. Subsequently, we obtain Figure 3 which illustrates the results.

Scenario 4: The parameters are set as $n = 3000$, $p = 8$, $\alpha = 0.05$. In this scenario, $ratio$ is varied to 0.02, 0.03, 0.04, 0.05, and 0.06. Subsequently, we obtain Figure 4 which illustrates the results.

B. Simulation analysis

To evaluate the prediction accuracy in data simulation, several metrics are considered: the MSE and MAE. The MSE and MAE formulas related to error are as follows

$$MSE(\hat{Y}) = \frac{1}{n}\sum_{i=1}^{n} |\hat{Y}_i - Y_i|^2,$$

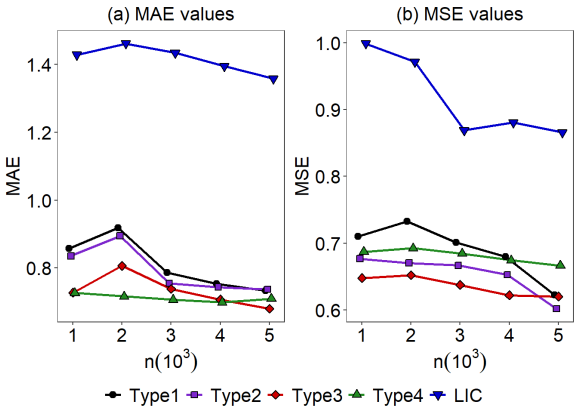

Fig. 1. The comparison results for Scenario 1 in Poisson distribution

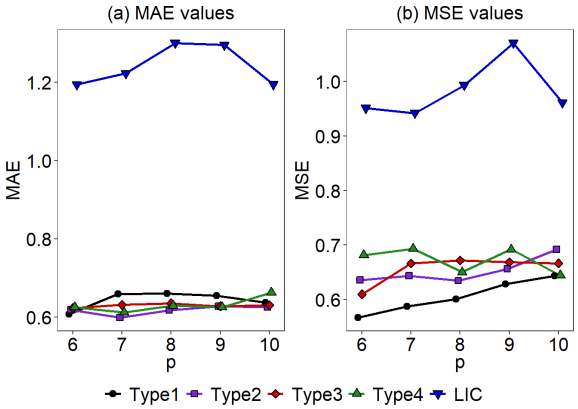

Fig. 2. The comparison results for Scenario 2 in Poisson distribution

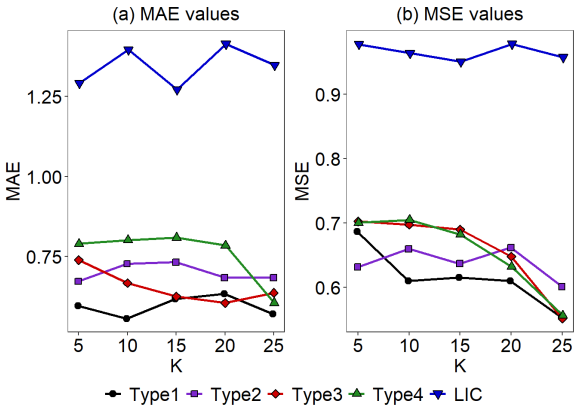

Fig. 3. The comparison results for Scenario 3 in Poisson distribution

$$MAE(\hat{Y}) = \frac{1}{n}\sum_{i=1}^{n} |\hat{Y}_i - Y_i|.$$

Case 1. Poisson Distribution

$$X_2 = (X_{2ij}) \in R^{m\times p}, X_{2ij} \sim Pois(\lambda).$$

Figure 1 indicate that the PPC outperforms the LIC notably when n equals 2000. Specifically, the MSE and MAE for the LIC are approximately 1.462 and 0.971, respectively, whereas for Type 4, these values are reduced to 0.716 and 0.692. Similarly, when p is 9, the PPC demonstrates significantly higher accuracy compared to the LIC. In this scenario, the MSE and MAE for the LIC are 1.413 and 0.978, respectively, whereas the MSE for Type 3 is 0.605 and the

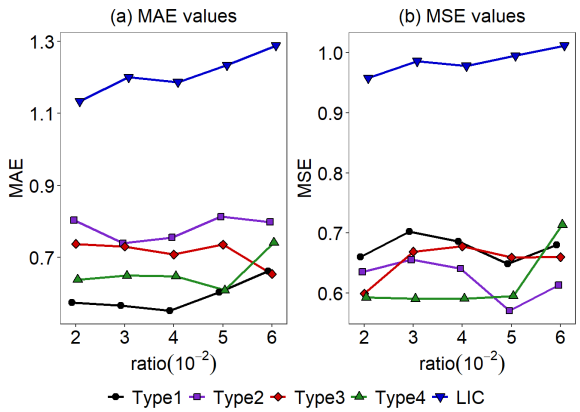

Fig. 4. The comparison results for Scenario 4 in Poisson distribution

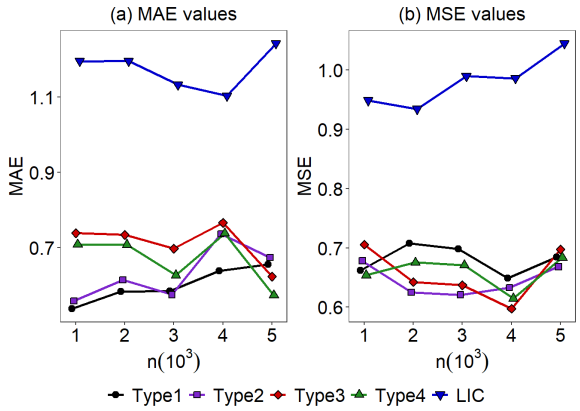

Fig. 5. The comparison results for Scenario 1 in Exponential distribution

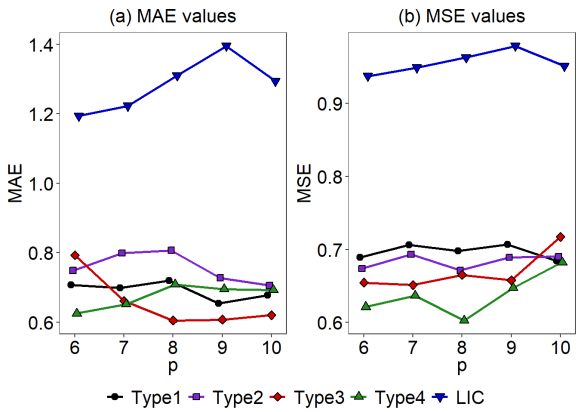

Fig. 6. The comparison results for Scenario 2 in Exponential distribution

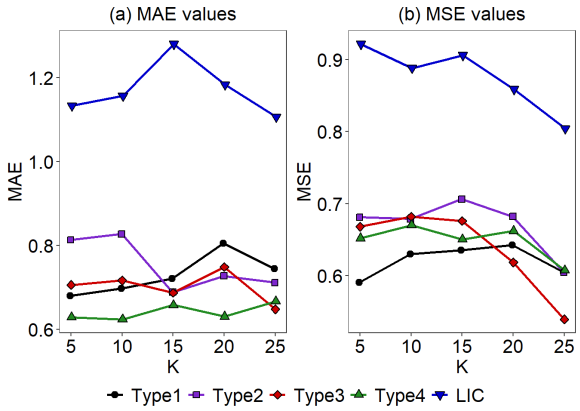

Fig. 7. The comparison results for Scenario 3 in Exponential distribution

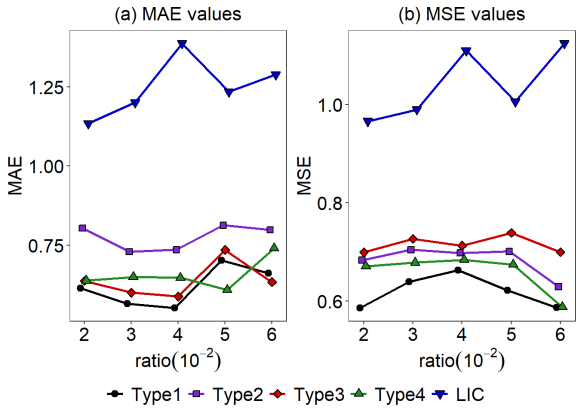

Fig. 8. The comparison results for Scenario 4 in Exponential distribution

MAE for Type 1 is 0.610. Analysis of Figure 3 suggests that the PPC exhibits greater stability than the LIC across varying values of K. Furthermore, Figure 4 reveals that when adjusting the proportion of redundant data, represented by $ratio$, the performance of the LIC is particularly poor when $ratio$ is 0.04.

A detailed analysis of Figures 3 and 4 reveals that the PPC shows more consistent performance than the LIC , especially under varying K. Notably, the performance of LIC significantly deteriorates at a redundancy level of 0.04 in Figure 4, indicating its limitation in handling highly redundant datasets.

These findings reinforce the notion that the PPC not only excels in precision but also maintains a higher degree of robustness and adaptability across varying levels of dataset complexity and redundancy.

Case 2. Exponential Distribution

$$X_2 = (X_{2ij}) \in R^{m\times p}, X_{2ij} \sim Exp(\theta).$$

At first, our task is to create a data set based on the exponential distribution, using a parameter value of $\theta = 5$.

This process allows us to generate a random dataset that accurately reflects specific parameters for our analysis.

Referring to Figure 5, it's evident that the PPC performs better than the LIC as the value of n increases. Specifically, when n reaches 5000, the LIC's MSE and MAE are approximately 1.242 and 1.044, respectively. The MSE of the PPC ranges from 0.60 to 0.80, while its MAE consistently remains below 0.80. This emphasizes that as n increases, the PPC demonstrates superior stability and accuracy compared to the LIC. When p is set to 8 or 9, the MSE and MAE of the LIC are relatively high, with the MSE peaking at 1.242. In contrast, the maximum MSE of the PPC is 1.044.

Figure 7 offers additional insight, revealing that while the error metrics of the LIC do decrease with an increase in K, the PPC sustains a higher level of stability and consistently outperforms with lower MSE and MAE values.

Lastly, Figure 8 explores variations in the proportion of redundant data denoted as $ratio$. The PPC solidify its position as a highly effective tool.

Case 3. Negative Binomial Distribution

$$X_2 = (X_{2ij}) \in R^{m\times p}, X_{2ij} \sim NB(\theta).$$

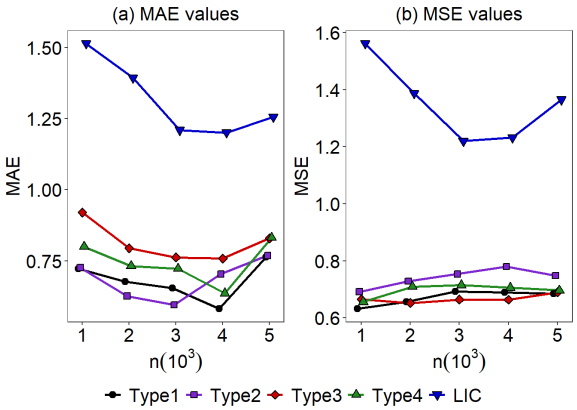

Fig. 9. The comparison results for Scenario 1 in Negative Binomial distribution

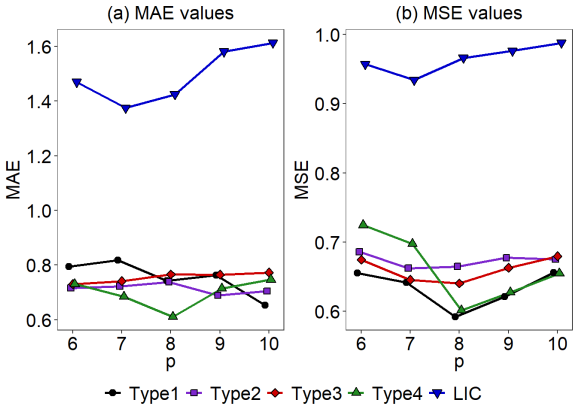

Fig. 10. The comparison results for Scenario 2 in Negative Binomial distribution

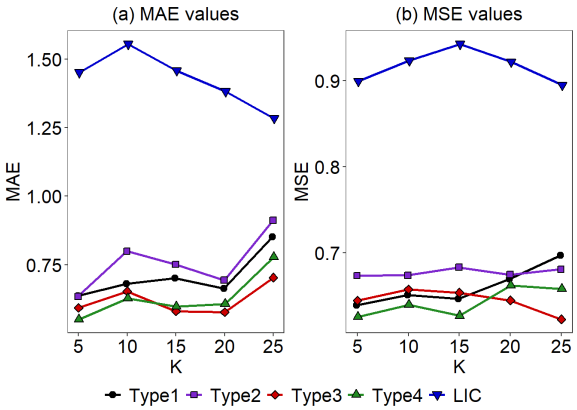

Fig. 11. The comparison results for Scenario 3 in Negative Binomial distribution

Our data set is generated by a foundation of the negative binomial distribution, meticulously calibrated with parameters $\gamma = 10$ and $\theta = 1$. This method fosters the synthesis of a randomized dataset meticulously tailored to conform to these exact specifications, thereby establishing a robust analytical baseline.

In Figure 9, the MSE and MAE related to the PPC show a strong clustering effect, indicating increased stability. This characteristic of the PPC becomes even more noticeable as the variable K changes. Additionally, when we adjust the $ratio$ parameter, the MSE and MAE for LIC being much

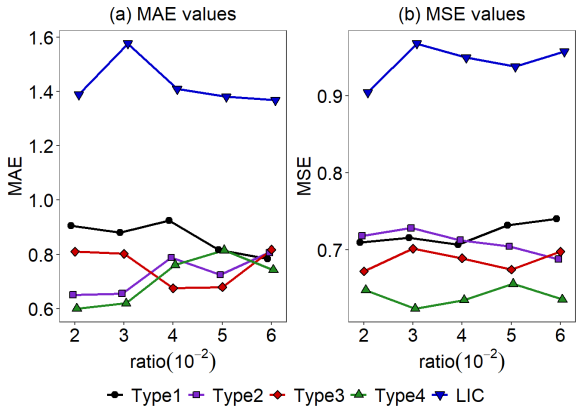

Fig. 12. The comparison results for Scenario 4 in Negative Binomial distribution

higher than those for the PPC.

IV. Conclusion

This paper delves into the examination of optimal subset selection for distributed interval estimation and hypothesis testing, utilizing fixed subset size processing as a key methodology. Through rigorous analysis, It becomes clear that the PPC method outperforms the LIC method, with reductions in MSE and MAE indicating a minimum improvement of 30%. This performance advantage of the PPC is further underscored by its notable stability across different values of K.

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

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

[4] 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.

[5] 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.

[6] P. Ma and X. Sun, "Leveraging for big data regression," *Computational Statistics*, vol. 7, no. 1, pp. 70-76, 2015.

[7] Yaqiong Yao and HaiYing Wang, "A Review on Optimal Subsampling Methods for Massive Datasets," *Journal of Data Science*, vol. 19, no. 1, pp. 151-172, 2021.

[8] Lih-Yuan Deng, Ching-Chi Yang, Dale Bowman, Dennis K. J. Lin, and Henry Horng-Shing Lu, "Big Data Model Building Using Dimension Reduction and Sample Selection," *Journal of Computational and Graphical Statistics*, vol. 33, pp. 435-447, 2023.

[9] Jun Yu, Jiaqi Liu, and Hai Ying Wang, "Information-based optimal subdata selection for non-linear models," *Statistical Papers*, vol. 64, no. 4, pp. 1069-1093, 2023.

[10] Q. Cheng, H. Y. Wang, and M. Yang, "Information-based optimal subdata selection for big data logistic regression," *Journal of Statistical Planning and Inference*, vol. 209, pp. 112-122, 2020.

[11] 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

[12] 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.

[13] 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.