Student LIC for Distributed Estimation

Guofu Jing, Guangbao Guo

***Abstract*—The Student t-linear regression method is widely used in statistical analysis. It proves particularly beneficial when dealing with a small sample or when the error distribution is irregular. The model is based on the ordinary least squares method and assumes that the random error term follows Student t distribution. This feature gives it robustness and flexibility, allowing it to provide more accurate parameter estimates and hypothesis testing results. In the simulation experiment section, we compared the proposed Student LIC criterion with two other indicators to verify its effectiveness. We also offer an in-depth exploration regarding the stability as well as the sensitivity of the LIC criterion under diverse redundant distributions. Experimental data show that the Student LIC criterion has excellent stability and can significantly reduce errors.**

***Index Terms*—T distribution, Student t-linear regression, LIC criterion, distributed estimation.**

I. Introduction

IN the field of statistical analysis, when the sample size of a dataset is limited or the error structure deviates from the normal, the difficulty of accurately estimating parameters increases. Traditional ordinary least squares (OLS) regression assumes that the error is normally distributed. However, in this case, this assumption is often insufficient and may lead to bias in the estimates. To solve this problem, this paper introduces the Student t regression model. The model uses a student t-distribution, which is characterized by a heavier tail. As a result, it provides a more flexible framework for datasets showing outliers or non-normal distribution. This integration enhances the model's robustness against outliers and provides it with greater flexibility.

A. Current research status

In the field of statistical analysis, research on redundant data has made significant progress. In this paper, we propose a new method depending on the Student t-linear regression model. This study is to solve the problem of estimating the distributed optimal subset in the context of redundant data. This method is especially suitable for datasets with small sample sizes or non-normal error distributions. It provides more reliable parameter estimation and hypothesis testing results.

Future research directions may include: improving existing methods and developing new algorithms to enhance the accuracy and stability of estimation under non-normal distribution and outlier conditions. Explore the potential application of this technology in fields that require handling complex data structures, such as financial risk management and precision medicine.

Manuscript received May 6, 2024; revised January 10, 2025.

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

Guofu Jing is an undergraduate student from Shandong University of Technology, Zibo, China. (e-mail:Guofu0806@163.com).

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

B. Our work

This paper deeply studies the theoretical characteristics of the Student LIC criterion. It also gives useful guidance for parameter selection and model evaluation in practical applications. We compared the MAE and MSE of three methods: the LIC criterion, the minimum information matrix, and the maximum gain matrix (respectively abbreviated as LIC, Iopt, and Lopt). The results confirm the criterion to be the optimal subset selection method. This paper also elaborates on the basic theory of the Student LIC criterion, and designs a series of simulation experiments. The purpose of these experiments is to select appropriate performance indicators. In addition, we explored the stability and sensitivity of these three methods in three common distribution functions.

II. Distributed student t-linear regression model

This section of the study focuses on student t regression model. It is expressed as:

$$Y_{I_k} = X_{I_k}\beta + \varepsilon_{I_k}, \quad \varepsilon_{I_k} \sim t(n_0), \quad k = 1, \ldots, K_n,$$

where n_0 denotes the degrees of freedom with $n_0 \geq 37$. X_{I_k} is a submatrix of size $n_{I_k} \times p$ with $n_{I_k} \geq p$, representing the vector of sub-residuals. The vector $\beta = (\beta_1, \ldots, \beta_p)^T$ contains the regression coefficients.

The total dataset can be represented in a matrix form:

$$Y = (Y_{I_1}^T, Y_{I_2}^T, \ldots, Y_{I_{K_n}}^T)^T, \quad X = (X_{I_1}^T, X_{I_2}^T, \ldots, X_{I_{K_n}}^T)^T,$$

thus, the model can be simply written as:

$$Y = X\beta + \varepsilon, \quad \varepsilon \sim t(n_0),$$

where Y is the random response variable and $X = (X_{ij})$ is the unknown matrix of size $n \times p$.

The concept of distributed estimation has gained significant momentum in this area. Specifically, a large amount of data from a single computer is distributed to a large number of computers, each of which generates local estimates by applying statistical inference methods. Subsequently, the local estimates are aggregated and averaged to produce a final estimate. In this case, the size of the block called K_n also plays a key role. If K_n is too large, one or more local estimation anomalies may occur, which will adversely affect the final result. Specifically, all data on a machine is randomly and evenly divided into K_n blocks, and each block is delivered to each machine in the following manner:

1)For $Y_{I_k} = (Y_{I_k,1}, \cdots, Y_{I_k,n_{I_k}})$, $k = 1, \ldots, K_n$, each machine computes the estimate $\hat{\mu}_{I_k}$ of the local mean $\mu = EY$ as follows:

$$\hat{\mu}_{I_k} = \bar{Y}_{I_k} = \sum_{i=1}^{n_{I_k}} Y_{I_k,i}/n_{I_k}.$$

2) Aggregate all local estimators and average them to get the final estimator:

$$\hat{\mu}^{(a)} = \frac{1}{K_n} \sum_{k=1}^{K_n} \hat{\mu}_{I_k},$$

where the variance of $\hat{\mu}_{I_k}$ is different, mean $var(\hat{\mu}_{I_i}) \neq var(\hat{\mu}_{I_j})(i \neq j)$.

$$\hat{\mu}_w = \sum_{k=1}^{K_n} w_k \hat{\mu}_{I_k}, w_k = \frac{\mu_{I_k}^{-1}}{\sum_{k=1}^{n_{I_k}} \mu_{I_k}^{-1}}, \mu_{I_k} = trace(var(\hat{\mu}_{I_k})),$$

where $var(\hat{\mu}_w) \leq var(\hat{\mu}^{(a)})$.

For the confidence interval $C(Y_{I_k})$ at a given confidence level of $1 - \alpha$, the probability is:

$$P(\mu_{I_k} \in C(Y_{I_k}) | \mu_{I_k}) = 1 - \alpha.$$

The function w, defined within the interval $w \in (0, 1)$, serves as the confidence domain function. The acceptance region $A_w(\mu_{I_k})$ for each μ_{I_k} is assumed to meet the following condition:

$$A_w(\mu_{I_k}) = \bar{Y}_{I_k} : (\mu_{I_k} - t_{n_{I_k}-p,1-\alpha w} \hat{\sigma}_{I_k} \cdot \bar{C}_{I_k}, \mu_{I_k} - t_{n_{I_k}-p,\alpha(1-w)} \hat{\sigma}_{I_k} \cdot \bar{C}_{I_k}),$$

where $\bar{C}_{I_k}$ is the average of C_{I_k,x_i} for all x_i in X_{I_k}, calculated as:

$$\bar{C}_{I_k} = \sum_{i=1}^{n_{I_k}} \frac{C_{I_k,x_i}}{n_{I_k}}, \quad x_i \in X_{I_k}$$

where C_{I_k,x_i} denotes the i-th diagonal element $X_{I_k}(X_{I_k}^T X_{I_k})^{-1} X_{I_k}^T$. Consequently, $\bar{C}_{I_k}$ can be expressed as the vector of these diagonal elements:

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

Specifically, the confidence interval is obtained by inverting the acceptance region at the α confidence level. When $w = \frac{1}{2}$, the confidence interval $C(Y_{I_k})$ for the mean μ_{I_k} is defined as:

$$C(Y_{I_k}) = \{\mu_{I_k} : \overline{Y}_{I_k} + t_{n_{I_k}-p,1-\frac{\alpha}{2}} \widehat{\sigma}_{I_k} \cdot \overline{C}_{I_k} \leq \mu_{I_k} \leq \overline{Y}_{I_k} + t_{n_{I_k}-p,1-\frac{\alpha}{2}} \widehat{\sigma}_{I_k} \cdot \overline{C}_{I_k}\},$$

where $E(\hat{\sigma}_{I_k}^2) = \sigma_{I_k}^2$. The calculation of is $\widehat{\sigma}_{I_k}^2$ given by:

$$\widehat{\sigma}_{I_k}^2 = \frac{1}{n_{I_k} - P} \hat{\varepsilon}_{I_k}^T \hat{\varepsilon}_{I_k} = \frac{1}{n_{I_k} - P} Y_{I_k}^T (I_{n_{I_k} \times n_{I_k}} - H_{I_k}) Y_{I_k},$$

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

For submatrices $X_{I_k}^T X_{I_k}$ of full rank, the matrix H_{I_k} is calculated as:

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

where λ represents the interference term, and $I_{n \times n}$ is the original matrix of $n \times n$.

Subsequently, the shortest interval length with respect to μ_{I_k} can be obtained:

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

III. Steps

i. **Optimal Subset Selection:** The LIC criterion is applied to determine the optimal subset I_{opt} through three steps:
   - First, determine I_{opt}^1 based on the shortest interval length.
   - Second, obtain I_{opt}^2 by maximizing the information matrix.
   - Finally, the intersection of I_{opt}^1 and I_{opt}^2 yields I_{opt}.

ii. **Simulation Preparation:** Generate simulated datasets using R software, ensuring that the error terms follow a Student t-distribution with specific degrees of freedom.

iii. **Error Calculation:** Calculate the MSE and MAE for the LIC criterion, as well as for I_{opt}^1 and I_{opt}^2 methods.

iv. **Performance Evaluation:** Compare the performance of LIC, I_{opt}^1, and I_{opt}^2 through line charts of MSE and MAE.

v. **Simulation Description and Analysis:** Describe the design of the simulation experiment and analyze the stability and sensitivity under different conditions.

vi. **Result Discussion:** Discuss the simulation results and assess the effectiveness of the LIC criterion in handling non-normal error distributions and outlier data.

IV. LIC criterion for optimal subset selection

In this study, we have meticulously designed a sequence of steps to select the optimal indicator subset sequence $\{I_k\}_{k=1}^{k_n}$, with the goal of enhancing estimation precision and reducing dataset size.

Step 1: The initial step is to identify the optimal indicator subset I_{opt}^1 based on the shortest interval length of μ_{I_k}, This is achieved by minimizing the expression:

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

where $\widehat{\sigma}_{I_k}, \bar{C}_{I_k}$ and $t_{n_{I_k}-1,1-\frac{\alpha}{2}}$ are derived from $\mathrm{L(C(Y}_{I_k})) = 2\widehat{\sigma}_{I_k} \cdot \bar{C}_{I_k} \cdot t_{n_{I_k}-p,1-\frac{\alpha}{2}}$.

Step 2: The LSE of β_{I_k} and the variance of $\hat{\beta}_{I_k}$ are then demonstrated as follows:

$$\hat{\beta}_{I_k} = \left(X_{I_k}^T X_{I_k}\right)^{-1} X_{I_k}^T Y_{I_k}, var\left(\hat{\beta}_{I_k}\right) = \hat{\sigma}_{I_k}^2 \left(X_{I_k}^T X_{I_k}\right)^{-1},$$

where $E(\widehat{\sigma}_{I_k}^2) = \sigma_{I_k}^2$. Building on this, the optimal indicator subset I_{opt}^2 is found by maximizing the information matrix $X_{I_k}^T X_{I_k}$:

$$I_{opt}^2 = \arg\max_{I_k} \left|X_{I_k}^T X_{I_k}\right|.$$

This step is similar to the IBOSS algorithm proposed by Wang Haiying under the D-op criterion. The algorithm selects a subset from K_n two-dimensional variables (Y_{I_k}, X_{I_k}), maximizing the formula:

$$\delta_{opt}^D = \arg\max_{\delta} \left|\sum_{k=1}^{K_n} \delta_k X_{I_k} X_{I_k}^T\right|, \sum_{k=1}^{K_n} \delta_k = i,$$

where δ_k represents the indicator variable. When $\delta_k = 1$, the subset includes (Y_{I_k}, X_{I_k}). Conversely, $\delta_k = 0$, the subset excludes (Y_{I_k}, X_{I_k}).

Step 3: To further eliminate redundant information and reduce the subset, the final optimal subset is calculated as:

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

Consequently, from all subsets $\{Q = (Y_{I_k}, X_{I_k})\}_{k=1}^{K_n}$, an optimal subset $Q_{I_{opt}} = (Y_{I_{opt}}, X_{I_{opt}})$ is derived. For this subset, the shortest interval length of $\mu_{I_{\text{opt}}}$ is given by:

$$L(C(Y_{I_{opt}})) = \widehat{\sigma}_{I_{opt}} \cdot \bar{C}_{I_{opt}} \cdot t_{n_{I_{opt}}-1,1-\frac{\alpha}{2}}.$$

V. Simulation study

In this part, the proposed LIC criterion's performance is evaluated by means of simulated data. Additionally, the performance of two other indicators, opt_1 and opt_2, is analyzed under identical conditions. The purpose is to conduct a comparative analysis. This analysis aims to elucidate the advantages of the LIC criterion over other indicators in a more comprehensive manner.

A. Simulation preparation

In this study, the performance of various indicator subsets was assessed by calculating the MSE and MAE of the estimates $\hat{\mu}$ based on I^1_{opt}, I^2_{opt}, and I_{opt}. Generally, lower values of these metrics indicate better prediction performance.

The estimates for each subset were defined as follows:

$$\hat{\mu}_{I^1_{opt}} = X_{I^1_{opt}}\hat{\beta}_{I^1_{opt}}, \hat{\mu}_{I^2_{opt}} = X_{I^2_{opt}}\hat{\beta}_{I^2_{opt}}, \hat{\mu}_{I_{opt}} = X_{I_{opt}}\hat{\beta}_{I_{opt}}.$$

For each subset estimate $\hat{\mu}_{I_k}$, the MSE is calculated as:

$$MSE(\hat{\mu}_{I_k}) = \frac{1}{n_{I_k}}[(Y_{I_k} - \widehat{Y}_{I_k})^T(Y_{I_k} - \widehat{Y}_{I_k})].$$

Using this definition, the MSE of the one-step average estimate $\hat{\mu}^{(a)}$ and the one-step median estimate $\hat{\mu}^{(m)}$ can be calculated as:

$$\begin{aligned} MSE(\hat{\mu}^{(a)}) = \min_k\{&\frac{1}{n_{I_k}}[(Y_{I_k} - \widehat{X}_{I_k}\hat{\beta}^{(a)})^T \\ &\times (Y_{I_k} - \widehat{X}_{I_k}\hat{\beta}^{(a)})]\} \\ MSE(\hat{\mu}^{(m)}) = \min_k\{&\frac{1}{n_{I_k}}[(Y_{I_k} - \widehat{X}_{I_k}\hat{\beta}^{(m)})^T \\ &\times (Y_{I_k} - \widehat{X}_{I_k}\hat{\beta}^{(m)})]\} \end{aligned}$$

For the specific subsets I^1_{opt} and I^2_{opt}, their MSE are calculated as:

$$MSE(\hat{\mu}_{I^1_{opt}}) = \frac{1}{n_{I^1_{opt}}}[(Y_{I^1_{opt}} - \widehat{Y}_{I^1_{opt}})^T(Y_{I^1_{opt}} - \widehat{Y}_{I^1_{opt}})],$$

$$MSE(\hat{\mu}_{I^2_{opt}}) = \frac{1}{n_{I^2_{opt}}}[(Y_{I^2_{opt}} - \widehat{Y}_{I^2_{opt}})^T(Y_{I^2_{opt}} - \widehat{Y}_{I^2_{opt}})].$$

The MSE for the optimal subset $\hat{\mu}_{I_{opt}}$ is calculated as:

$$MSE(\hat{\mu}_{I_{opt}}) = \frac{1}{n_{I_{opt}}}(Y_{I_{opt}} - \hat{Y}_{I_{opt}})^T(Y_{I_{opt}} - \hat{Y}_{I_{opt}}).$$

The MAE is defined for the five estimates:

$$\begin{aligned} MAE(\hat{\mu}^{(a)}) &= \min_k\{|\overline{Y}_{I_k} - \hat{\mu}^{(a)}|\}, \\ MAE(\hat{\mu}^{(m)}) &= \min_k\{|\overline{Y}_{I_k} - \hat{\mu}^{(m)}|\}, \\ MAE(\hat{\mu}_{I^1_{opt}}) &= |\overline{Y}_{I_k} - \hat{\mu}_{I^1_{opt}}|, \\ MAE(\hat{\mu}_{I^2_{opt}}) &= |\overline{Y}_{I_k} - \hat{\mu}_{I^2_{opt}}|, \\ MAE(\hat{\mu}_{I_{opt}}) &= |\overline{Y}_{I_k} - \hat{\mu}_{I_{opt}}|. \end{aligned}$$

B. Stability analysis

I: Simulation Description

The focus of this section is to explore the stability of the LIC criterion under different conditions. Specifically, we assume that the error term follows Student t-distribution. At the same time, the generation process of the dataset X_2 follows other different distributions, including uniform distribution, chi-squared distribution, and geometric distribution. Under these conditions, we conduct an analysis of the stability of the LIC criterion.

The dataset (X, Y) is generated as follows:

$$Y_i = X_i\beta + \varepsilon_i, \varepsilon_i \sim t(n_0),$$

where X is composed of (X_1, X_2) while Y is made up of (Y_1, Y_2). The definitions are as follows:

$$\begin{aligned} X_1 = (X_{ij}) \in IR^{n_1\times p}, X_{1ij} \sim N(0,2), \\ X_2 = (X_{ij}) \in IR^{n_2\times p}, X_{2ij} \sim, \\ Y_1 = X_1\beta + \varepsilon_1, n_1 = n - n_r, \\ Y_2 = X_2\beta + \varepsilon_2, n_2 = n_r. \end{aligned}$$

The cases differ based on the distribution of X_{2ij}:

1) Case 1 (Uniform Distribution): $X_{2ij} \sim \text{Unif}(0, 3)$
2) Case 2 (Chi-Square Distribution): $X_{2ij} \sim \chi^2(20)$
3) Case 3 (Geometric Distribution): $X_{2ij} \sim \text{Geom}(0.6)$

It is noted that $\beta \sim Unif(0, 3)$ and $\varepsilon = (\varepsilon_1, \varepsilon_2)$ where $\varepsilon_1 \sim$t$(exp(exp(0.5 - X_2)))$, $\varepsilon_2 \sim$t$(exp(exp(0.5 - (X_2))))$, and then run our simulation.

This section studies the stability of the LIC criterion under different distributions by changing the values of n and p.

II: Simulation Analysis

Case1. This case study is the stability of the LIC criterion under uniform distribution conditions.

i: The impact of n-value on the stability of the LIC

In this experiment, the settings for the values are as follows: $p = 8$, $K = 10$, $\alpha = 0.01$, $\sigma_1 = 3$, $\sigma_2 = 5$, and $n_r = 50$. Under these conditions, the sample size n varies within the set $\{1000, 2000, 3000, 4000, 5000\}$.

Fig. 1. shows that under uniform distribution, both MAE and MSE of the LIC method showed good stability as the sample size n increased. Specifically, MAE fluctuates greatly when the sample size increases from 1000 to 2000, but its value gradually stabilizes and decreases significantly in the range of n reaching 2000 to 5000. At the same time, MSE has remained at a low level with minimal variation. These results show that the LIC can effectively control the error and maintain high stability under different sample sizes.

ii: The impact of p-value on the stability of the LIC

In this experiment, the settings for the values are as follows: $n = 2000$, $K = 10$, $\alpha = 0.01$, $\sigma_1 = 3$, $\sigma_2 = 5$, and $n_r = 50$. Under these conditions, the sample size p varies within the set $\{8, 9, 10, 11, 12\}$.

Fig. 2. shows that under uniform distribution, as the p-value increases from 8 to 12, the LIC method exhibits higher stability and lower error at all p-values. In contrast, the MAE and MSE of the Opt1 and Opt2 methods fluctuate more significantly at different p-values.

Case2. This case study is the stability of the LIC criterion under chi-square distribution conditions.

i: The impact of n-value on the stability of the LIC

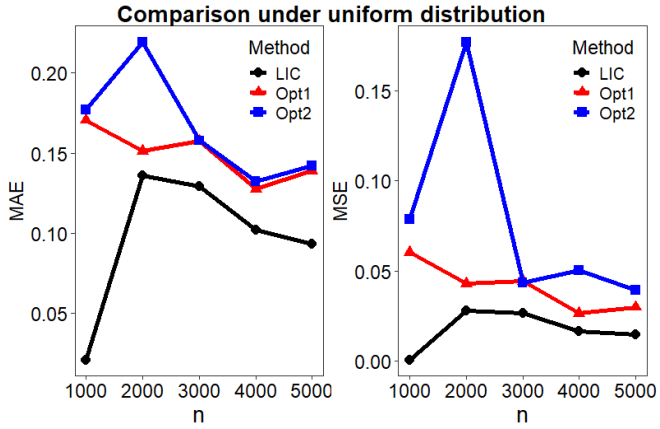

Fig. 1. Stability analysis of LIC for n-value variations under uniform distribution.

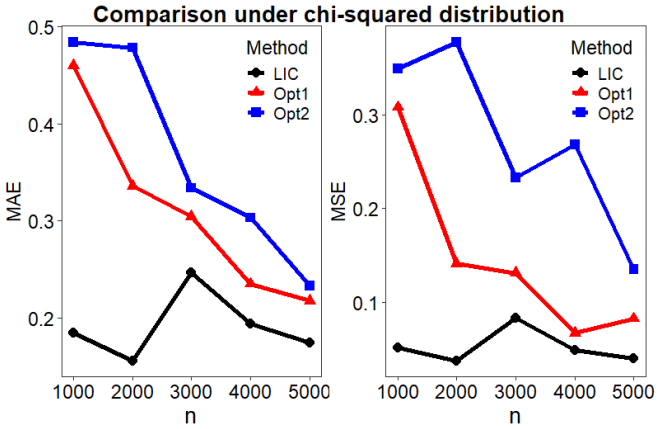

Fig. 3. Stability analysis of LIC for n-value variations under chi-squared distribution.

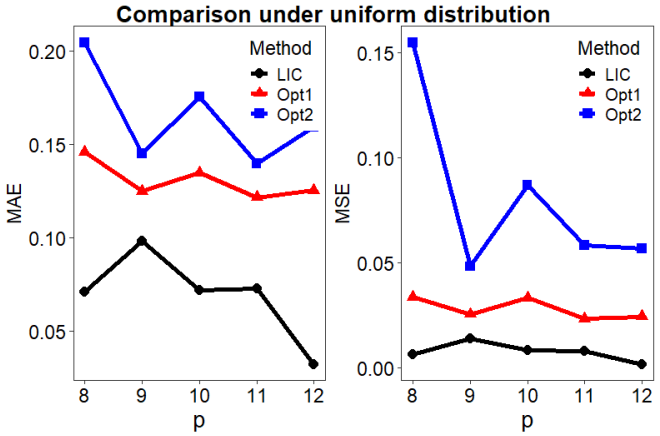

Fig. 2. Stability analysis of LIC for p-value variations under uniform distribution.

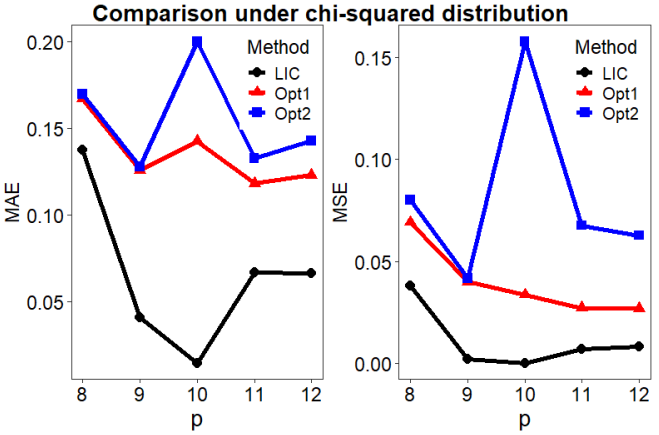

Fig. 4. Stability analysis of LIC for p-value variations under chi-squared distribution.

In this experiment, the settings for the values are as follows: $p = 8$, $K = 10$, $\alpha = 0.01$, $\sigma_1 = 3$, $\sigma_2 = 5$, and $n_r = 50$. Under these conditions, the sample size n varies within the set $\{1000, 2000, 3000, 4000, 5000\}$.

Fig. 3. shows that under the chi-square distribution, the MAE and MSE of all methods show a decreasing trend. This indicates that as the amount of data increases, the error of the model decreases. However, the LIC method shows the lowest MAE and MSE under all sample sizes. In contrast, the error metrics of the Opt1 and Opt2 methods fluctuate greatly with different sample sizes.

ii: The impact of p-value on the stability of the LIC

In this experiment, the settings for the values are as follows: $n = 2000$, $K = 10$, $\alpha = 0.01$, $\sigma_1 = 3$, $\sigma_2 = 5$, and $n_r = 50$. Under these conditions, the sample size p varies within the set $\{8, 9, 10, 11, 12\}$.

Fig. 4. shows that under the chi-square distribution, the LIC method shows lower values on both MAE and MSE metrics. This may indicate that LIC has better adaptability to variable dimension changes under the chi-square distribution. The Opt1 and Opt2 methods show greater fluctuations on MAE and MSE. Especially at $p = 10$, the fluctuation of Opt2 is particularly significant. The results show that the LIC method may have potential advantages in controlling errors under the condition of chi-square distribution.

Case3. This case study is the stability of the LIC criterion under geometric distribution conditions.

i: The impact of n-value on the stability of the LIC

In this experiment, the settings for the values are as follows: $p = 8$, $K = 10$, $\alpha = 0.01$, $\sigma_1 = 3$, $\sigma_2 = 5$, and $n_r = 50$. Under these conditions, the sample size n varies within the set $\{1000, 2000, 3000, 4000, 5000\}$.

Fig. 5. shows that under the geometric distribution, as the sample size n increases, the LIC method shows lower and stable values on both the MAE and MSE key error indicators. This implies that its good robustness in handling geometric distributed data. In contrast, the Opt1 and Opt2 methods, although also showed a tendency for errors to decrease with increasing sample size. However, the error indicators of the Opt1 and Opt2 methods fluctuate greatly under different sample sizes.

ii: The impact of p-value on the stability of the LIC

In this experiment, the settings for the values are as follows: $n = 2000$, $K = 10$, $\alpha = 0.01$, $\sigma_1 = 3$, $\sigma_2 = 5$, and $n_r = 50$. Under these conditions, the sample size p varies within the set $\{8, 9, 10, 11, 12\}$.

Fig. 6. shows that under the geometric distribution, the MAE and MSE values of the LIC method always remain the lowest among all p-values. In addition, the MSE curve of the LIC method shows high stability and less fluctuation when p changes. In contrast, the values of the Opt1 and Opt2 methods fluctuate significantly under different p-values.

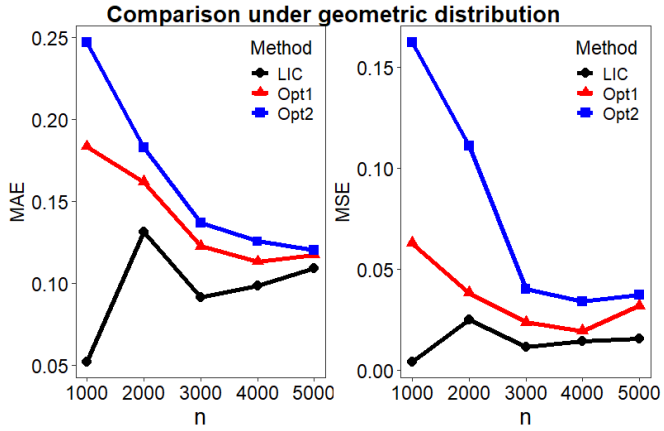

Fig. 5. Stability analysis of LIC for n-value variations under geometric distribution.

Especially at $p = 9$, the MSE value of the Opt2 method appears a significant peak. Although Opt1 and Opt2 also show certain stability under some p-values. Overall, the LIC method performs better in controlling errors and maintaining stability.

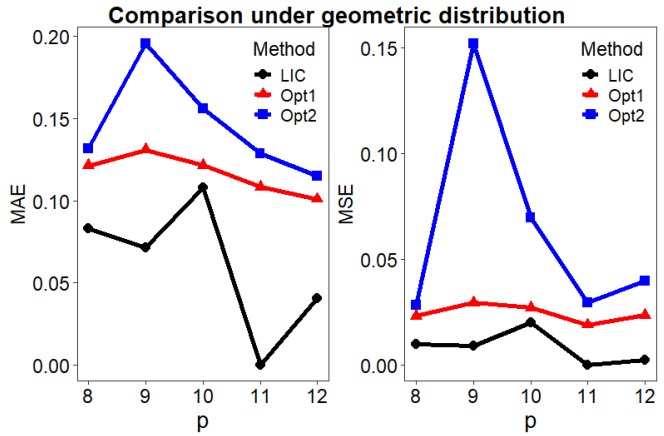

Fig. 6. Stability analysis of LIC for p-value variations under geometric distribution.

C. Sensitivity analysis

This section examines the sensitivity of the LIC principle under uniform, chi-square, and geometric distributions by changing the values of K and n_r. The simulation description in this section is the same as the simulation description in the stability analysis, so it will not be repeated here.

I: Simulation Analysis

Case 4. This case study is the sensitivity of the LIC criterion under uniform distribution conditions.

i: The impact of K-value on the sensitivity of the LIC

In this experiment, the settings for the values are as follows: $p = 8$, $n = 6000$, $\alpha = 0.01$, $\sigma_1 = 3$, $\sigma_2 = 5$, and $n_r = 50$. Under these conditions, the sample size K varies within the set $\{5, 10, 15, 20, 25\}$.

Fig. 7. shows that under the uniform distribution, the MAE of the LIC method fluctuates within a smaller range under different K values. In contrast, the MAE of the Opt1 and Opt2 methods fluctuate greatly. At the same time, for MSE, the LIC method maintains a very low level over the entire K-value range, almost close to zero, and there is no significant fluctuation. From the trend analysis, MAE and MSE of LIC method have relatively stable change trend under different K values, but Opt1 and Opt2 methods have obvious change trend.

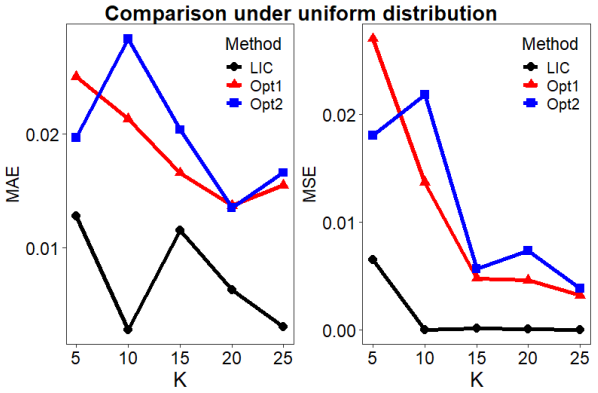

Fig. 7. Sensitivity analysis of LIC for K-value variations under uniform distribution.

ii: The impact of n_r-value on the sensitivity of the LIC

In this experiment, the settings for the values are as follows: $p = 8$, $n = 2000$, $\alpha = 0.01$, $\sigma_1 = 3$, $\sigma_2 = 5$, and $n_r = 50$. Under these conditions, the sample size n_r varies within the set $\{50, 60, 70, 80, 90\}$.

Fig. 8. shows that under the uniform distribution, the LIC method shows some sensitivity to changes in n_r on both the MAE and MSE error metrics, but the overall error level remains in the low range. In particular, the MAE and MSE of the LIC method fluctuated slightly during the increase of the n_r value from 50 to 90, but the magnitude of the change was small compared to that of the Opt1 and Opt2 methods.

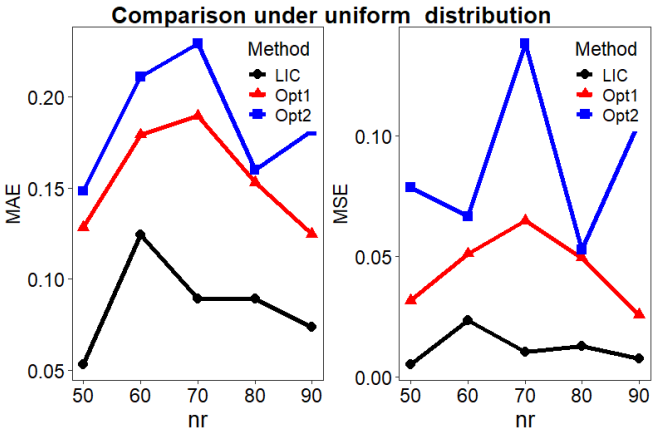

Fig. 8. Sensitivity analysis of LIC for n_r-value variations under uniform distribution.

Case 5. This case study is the sensitivity of the LIC criterion under chi-square distribution conditions.

i: The impact of K-value on the sensitivity of the LIC

In this experiment, the settings for the values are as follows: $p = 8$, $n = 6000$, $\alpha = 0.01$, $\sigma_1 = 3$, $\sigma_2 = 5$, and $n_r = 50$. Under these conditions, the sample size K varies within the set $\{5, 10, 15, 20, 25\}$.

Fig. 9. shows that under the chi-square distribution, the fluctuation of LIC is smaller compared to Opt1 and Opt2. As the K value increases from 5 to 25, the MAE of LIC changes more smoothly. In particular, LIC also shows low sensitivity in terms of MSE. Its MSE values fluctuate very little throughout the K-value range and remain consistently low. However, the MSE values of Opt1 and Opt2 fluctuate more. Overall, the LIC method is less sensitive to the change of K value under the chi-square distribution and has better stability.

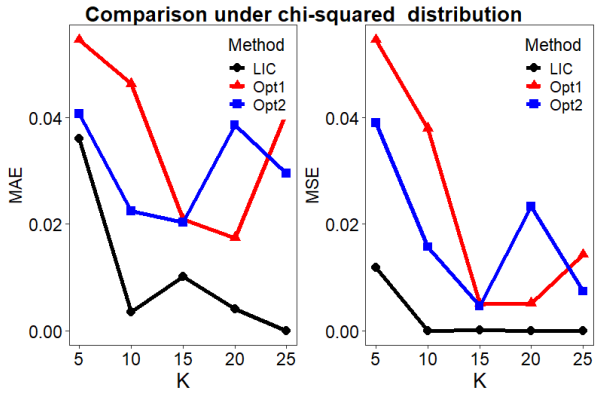

Fig. 9. Sensitivity analysis of LIC for K-value variations under chi-square distribution.

ii: The impact of n_r-value on the sensitivity of the LIC

In this experiment, the settings for the values are as follows: $p = 8$, $n = 2000$, $\alpha = 0.01$, $\sigma_1 = 3$, $\sigma_2 = 5$, and $n_r = 50$. Under these conditions, the sample size n_r varies within the set $\{50, 60, 70, 80, 90\}$.

Fig. 10. shows that under the chi-square distribution, as the value of n_r changes, the MAE and MSE curves of the LIC method show different trends compared to Opt1 and Opt2. The MAE of the LIC reaches a local peak at $n_r = 60$, and then decreases with the increase of n_r. Similarly, the MSE of LIC also reaches a peak at $n_r = 60$, and then shows a decreasing trend as n_r continues to increase. These fluctuations may indicate that the LIC method is sensitive to changes in the parameter n_r, especially when the n_r value is low.

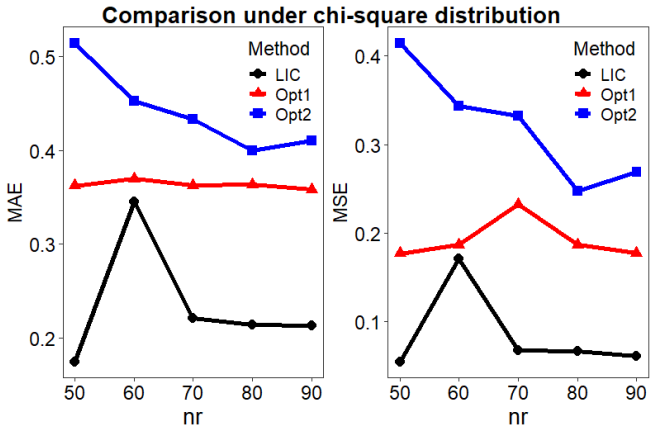

Fig. 10. Sensitivity analysis of LIC for n_r-value variations under chi-square distribution.

Case 6. This case study is the sensitivity of the LIC criterion under geometric distribution conditions.

i: The impact of K-value on the sensitivity of the LIC

In this experiment, the settings for the values are as follows: $p = 8$, $n = 6000$, $\alpha = 0.01$, $\sigma_1 = 3$, $\sigma_2 = 5$, and $n_r = 50$. Under these conditions, the sample size K varies within the set $\{5, 10, 15, 20, 25\}$.

Fig. 11. shows that under the geometric distribution, the MAE value of the LIC method remains at a low level and fluctuates little when the K value changes. Moreover, its MSE value also remains the lowest throughout the K value range, with a small change in amplitude. This indicates that the LIC method can maintain its performance consistency well in the face of changes in the parameter K. In contrast, the Opt1 and Opt2 methods show greater fluctuations when certain K values change.

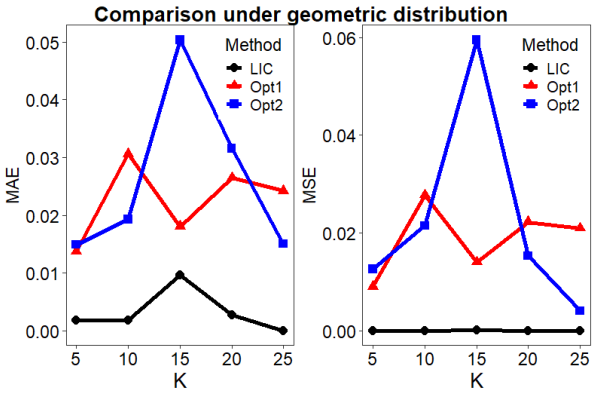

Fig. 11. Sensitivity analysis of LIC for K-value variations under geometric distribution.

ii: The impact of n_r-value on the sensitivity of the LIC

In this experiment, the settings for the values are as follows: $p = 8$, $n = 2000$, $\alpha = 0.01$, $\sigma_1 = 3$, $\sigma_2 = 5$, and $n_r = 50$. Under these conditions, the sample size n_r varies within the set $\{50, 60, 70, 80, 90\}$.

Fig. 12. shows that under the geometric distribution, the LIC method shows a certain sensitivity to the change of parameter n_r. On the two error indicators of MAE and MSE, the error value of the LIC method fluctuates with the change of n_r, especially around $n_r = 80$, the MAE value of the LIC method shows a local peak. And the MSE values remained low and stable over the entire nr range. This sensitivity indicates that the performance of the LIC method may be affected to a certain extent when adapting to the change of the n_r parameter, but overall, it can still maintain a low error level.

D. Summary of the simulation

In the simulation experiment of numerical analysis, the stability and sensitivity when dealing with different distribution data were deeply explored. The experimental results show that the criterion exhibits strong stability and sensitivity. It can effectively reduce errors and improve the accuracy and reliability of data. Especially when the data size and dimension change, the LIC criterion shows good stability. It can maintain a low error value and a relatively stable curve trend. Most importantly, these findings highlight the superiority of the LIC criterion in handling different data distributions. They provide theoretical support for the further research and application of this criterion.

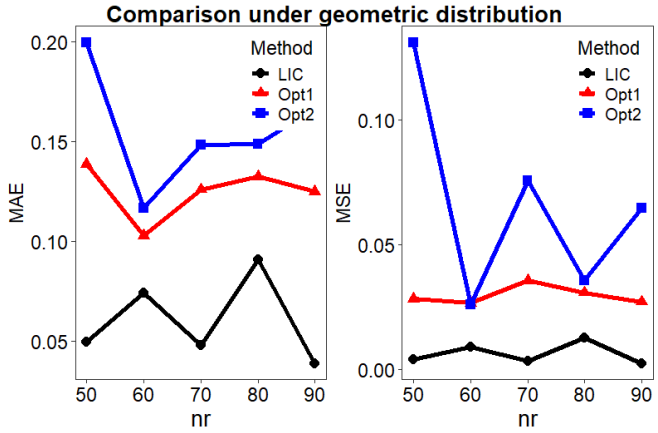

Fig. 12. Sensitivity analysis of LIC for n_r-value variations under geometric distribution.

This study verified the effectiveness of the Student LIC standard through simulation experiments. Although the research results are quite insightful, we must admit that the generalizability of these results may be limited by sample selection and model assumptions. For example, if the data distribution deviates from our assumed distribution, the accuracy of the model will decrease. Future research should test the stability and reliability of the LIC standard under a wider range of conditions.

VI. Conclusion

This paper discusses the theoretical basis of Student t-linear regression model. And discussed its applicability in distributed data estimation. In particular, the comparison between LIC and other subset selection methods highlights the superiority of the Student LIC principle in dealing with non-normal error distribution and outlier data. The research results show that the Student LIC criterion can not only achieve optimal subset selection, it can also effectively reduce redundant information and maintain a small credibility interval, thereby improving the estimation accuracy.

Data Availability

The criterion has been implemented by us in an R package called LIC, and this package has been publicly released. For more details, please visit the website at https://CRAN.R-project.org/package=LIC.

References

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

[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] J. Lederer. Fundamentals of High-Dimension Statistics. Switzerland: Springer Nature Switzerland AG, 2020. https://doi.org/10.1007/978-3-030-73792-4$_4$.

[6] G. Guo, Y. Sun, G. Qian, Q. Wang. LIC criterion for optimal subset selection in distributed interval estimation. *Journal of Applied Statistics*, vol. 50, no. 9, pp. 1900-1920, 2022. https://doi.org/10.1080/02664763.2022.2053949.

[7] G. Guo, Y. Sun, G. Qian, and Q. Wang. LIC: The LIC Criterion for Optimal Subset Selection. 2022. https://CRAN.R-project.org/package=LIC.

[8] T. C. Fonseca, M. A. Ferreira, H. S. Migon. Objective analysis for the Student-t regression model. *Biometrika*, vol. 95, no. 2, pp. 325–333, 2008.

[9] D. He, D. Sun, L. He. Objective analysis for the Student-t linear regression. *Analysis*, vol. 16, no. 1, pp. 129–145, 2021.

[10] C. Fernández, M. F. Steel. Multivariate Student regression models: Pitfalls and inference. *Biometrika*, vol. 86, no. 1, pp. 153–167, 1999.

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

[12] 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.

[13] G. Guo, M. Yu, and G. Qian. ORKM: Online regularized K-means clustering for online multi-view data. *Information Sciences*, vol. 680, p. 121133, 2024.

[14] 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, 2024.

[15] D. Chang, G. B. Guo. Research on Distributed Redundant Data Estimation Based on LIC. *IAENG International Journal of Applied Mathematics*, vol. 55, no. 1, pp. 1–6, 2025.

[16] J. Li, G. B. Guo. An Optimal Subset Selection Algorithm for Distributed Hypothesis Test. *IAENG International Journal of Applied Mathematics*, vol. 54, no. 12, pp. 2811–2815, 2024.

[17] Y. Li, G. B. Guo. General Unilateral Loading Estimation. *Engineering Letters*, vol. 32, no. 1, pp. 72–76, 2024.

[18] W. You, Z. Yang, G. B. Guo, X.-F. Wan, G. Ji. Prediction of DNA-binding proteins by interaction fusion feature representation and selective ensemble. *Knowledge-Based Systems*, vol. 163, pp. 598–610, 2018.

[19] W. Shao, G. B. Guo. Multiple-try simulated annealing algorithm for global optimization. *Mathematical Problems in Engineering*, vol. 2018, no. 1, pp. 1–11, 2018.

[20] W. Shao, G. B. Guo, G. Zhao, F. Meng. Simulated annealing for the bounds of Kendall's and Spearman's. *Journal of Statistical Computation and Simulation*, vol. 84, no. 12, pp. 2688–2699, 2014.

[21] W. Shao, G. B. Guo, F. Meng, S. Jia. An efficient proposal distribution for Metropolis–Hastings using a-splines technique. *Computational Statistics and Data Analysis*, vol. 57, pp. 465–478, 2012.

[22] G. B. Guo, S. Lin. Schwarz Method for Penalized Quasilikelihood in Generalized Additive Models. *Commun. Statist.-Theory Meth.*, vol. 39, pp. 1847–1854, 2010.