Distributed Data Analysis Based on Single Index Model

Jingcheng Xian, Cheng Wang and Guangbao Guo

***Abstract*—Amid randomized clinical trial data analysis, this article propose a distributed data analysis approach based on a single-index model that uniquely estimates the interaction between pre-processing covariates and treatment variables on the response variable. The method represents the interaction effects of the model via a set of therapy-specific adaptive link functions that act on a linear mixture of covariates (i.e., a single index) while satisfying the limitation that the expected value of the covariates is zero, while the primary effects of the covariates remain unspecified. By uniquely estimating the interaction effects between pre-processing covariates and treatment variables, we can optimize personalized treatment rules to improve clinical treatment outcomes.**

***Index Terms*—distributed single-index model, treatment optimization case data**

I. Introduction

A. Our Work

The research in this paper uses a single index model for distributed data analysis. This method spreads data over many computing nodes and uses one index to describe and analyze the data. This helps the method work well in distributed settings and be more efficient and accurate. The main parts of this paper are:

To do distributed data processing and analysis, we need to build a good and reliable distributed computing framework. This paper uses a constrained least squares method to improve the working model. We use a cubic spline estimate for the model, with constant weights for the modified covariate in the sample. In the end, we use penalized additive cubic splines least squares estimation to estimate additive regressions for each treatment.

To check if the distributed data analysis method using a single index model works well, we use random trial data for numerical analysis. When we compare it, the single index model shows big advantages. This method solves the efficiency and accuracy problems of multi-index models in a distributed environment and stays highly efficient and accurate.

In short, this paper looks into a distributed data analysis method using a single index model. By building a distributed computing framework, designing a single index model, and creating distributed data analysis algorithms, we can process and analyze data efficiently and accurately. Through experiments and analyzing the results, we show that this method works well and talk about future research directions.

Manuscript received July 12, 2024; revised May 18, 2025.

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.

Jingcheng Xian is an undergraduate student at Shandong University of Technology, Zibo, China (e-mail: xianjc0602@163.com).

Chen Wang is a graduate from Shandong University of Technology, Zibo, China. (e-mail: chenw0808@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).

II. Distributed Single Index Model

A. Distributed Single index Model

Amid big data, we consider preprocessing the covariates $X = (X_{I_1}, \ldots, X_{I_K})$. At the Kth node, X_{I_K} satisfies $X_{I_K} \in R^P$ and a discrete treatment variable $T \in \{1, \ldots, L\}$ with L categories, which has associated randomization probabilities $\{\pi_1, \ldots, \pi_L\}$.Let $Y = (Y_{I_1}, \ldots, Y_{I_K})$. For example, $Y_{I_K}^{(t)} \in R$ represents the potential outcome if a patient receives treatment $T = t$. We only measure $Y = Y^T$, T and X. In this paper, we presume that $E[Y_{I_K} \mid T = t] = 0$, the primacy effect of T, is mean-zero. This is done only to remove treatment-specific intercepts in the regression model to make it simpler to explain. It can be achieved by subtracting the treatment-specific mean of Y from t, and X is mean-zero.

This study mainly focuses on modeling the interaction effects between X and Y. We assume that $Y_{I_K} = E[Y_{I_K} \mid X_{I_K}, T] + \epsilon$, where ϵ represents zero-mean independent disturbance with finite variance. We Presume that the nested mean model linked to the interaction has a single-index framework. This includes a set of therapy-specific linking functions t for the single-index coefficient $a_0 \in R^P$, for $t = 1, ..., L$.

$$\mathrm{E}[\mathrm{Y}_{\mathrm{I_K}} \mid \mathrm{X}_{\mathrm{I_K}}, \mathrm{T} = \mathrm{t}] = \underbrace{\mu(X_{I_K})}_{\text{maineffect}} + \underbrace{\{f_t(\alpha_0^T X_{I_K})\}}_{\text{interaction}} \tag{1}$$

$\mu(X_{I_K})$ represents the main effect of X_{I_K}. In model (1), the treatment-specific function $f_t(\cdot)$ for t is a smooth univariate function.

To make the model identifiable without losing generality, it is assumed that the treatment-specific functions $f_i (i = 1...L)$ in model (1) meet a certain condition for treatment.

$$E\left[f_t\left(\alpha_0^T X_{I_K}\right) \mid X_{I_K}\right] = \sum_{t=1}^{L} \pi_t f_t\left(\alpha_0^T X_{I_K}\right) = 0 \tag{2}$$

The condition usually means that among the L interaction functions $(f_1, \ldots, f_L)$, only L-1 functions are free. In other words, in model (1), the last function f_L is determined by the other (L-1) functions, except for f_t.

$$f_t\left(\alpha_0^T X_{I_K}\right) = -\pi_L^{-1} \sum_{t=1}^{L-1} \pi_t f_t\left(\alpha_0^T X_{I_K}\right)$$

is almost inevitable. In model (1), because the linking function $f_t\,(t=1,\ldots,L)$ is nonparametric, the single-index coefficients α_0 can only be identified up to scale and sign. Therefore, without losing generality, we assume $\alpha_0 \in \Theta$, where

$$\Theta\colon = \{\alpha = (\alpha_1, ..., \alpha_p)^T \in R^P$$
$$:||\alpha|| = 1, \alpha_1 > 0\}$$

The semi-parametric model (1) characterizes the variability of X associated with treatment effects through a single index $\alpha_0^T X_{I_K} \in R$, and captures its interaction with treatment through a treatment-specific linking function $(f_1,\ldots,f_L)$. The interaction effects are driven by the different shapes of the unspecified function $f_i(i=1...L)$. There are multiple reasons why we examine a single index $\alpha_0^T X_{I_K}$ in model (1) instead of L indices specific to treatment. First, a general single index provides a concise one-dimensional integrated treatment effect modifier (expressed as a linear mixture of X) that allows for intuitive visualization of the interaction effects. In addition to its simplicity, the one-dimensional simplification in model (1) naturally generalizes linear model-based approaches. If L = 4 or L = 5 , we constrain the unspecified interaction function f in model (1) to pre-specified linear forms.

$$f_t\left(\alpha_0^T X_{I_K}\right) = (t + \pi_1 - 2)\,\alpha_0^T X_{I_K}$$

Then simplify the semi-paraindex model (1) into a modified covariate model.

To estimate the interaction term $f_t\left(\alpha_0^T X_{I_K}\right)$ for $t = 1,\ldots,L$ in model (1), given the unspecified main effects $\mu\left(X_{I_K}\right)$, we suggest using a working model.

For $\alpha \in \Theta$, we have the constraint:

$$\left[\mathbf{g}_T(\alpha^T X_{l_K}) \mid X_{l_K}\right] = \sum_{t=1}^{L} \pi_t \mathbf{g}_t(\alpha^T X_{l_K}) = 0 \quad (3)$$

It is likely that for all α , constraint (4) is applied to the smoothing link function $(g_1,\ldots,g_L)$ specific to the treatment t in the working model (3).

Within the least squares structure of model (3).

To fit the constrained operative model (3), we use a restricted least squares criterion.

$$I_{opt} = \underbrace{argmin}_{k=1...K} \quad E\left[\left(Y_{l_k} - g_T(\alpha^T X_{I_K})\right)^2/2\right] \quad (6)$$

We compute the minimum over k and find a subset I_K such that the elements in I_K are $1,\ldots,n$ forming a subset of $1,\ldots,n$ with n_{I_K} elements. In equation (6), the search for the minimum among K is called the extreme statistic X_{opt} and Y_{opt}.

B. Extreme Statistic

When comparing the effects of drug treatment and placebo treatment using extreme values as extreme statistics, follow these steps: Calculate the maximum pain level for the drug treatment group and the placebo treatment group. Let Y be the maximum pain level for the drug treatment group, and Y_{opt} be the maximum pain level for the placebo treatment group. Here, K is the individual index, and K_{opt} is the optimal K.

Assume the extreme value distribution function models the maximum values in each group. In this case, use the Gumbel distribution to fit the distribution of the extreme values. In drug treatment group, estimate the parameters (μ_1,β_1) of the Gumbel distribution. For the placebo treatment group, estimate the parameters (μ_2,β_2) of the Gumbel distribution.

$$f(y,\mu,\beta) = (1-\beta)\exp\Big((y-\mu)/\beta\Big)$$
$$* \exp\Big(-\exp\Big((y-\mu)/\beta\Big)\Big) \quad (7)$$

In drug treatment group, the predicted maximum value, denoted as X_{opt}, is generated based on the Gumbel distribution parameters (μ_1,β_1). For the placebo treatment group, the predicted maximum value, also denoted as X_{opt}, is generated based on the Gumbel distribution parameters (μ_2,β_2). Calculate the mean squared error between the predicted values and the actual observed values (the maximum pain level in each group), which is the average of the squared differences. Here, n_{opt} represents the sample size of both the drug treatment group and the placebo treatment group.

III. Distributed Simulation

A. Distributed Simulation Study of $L = 4$ Treatment Levels

In this section, we present additional simulation results to evaluate the performance of the restricted single-index model in assessing optimal treatment prediction policies when the number of treatment options L = 4. We consider different scenarios with varying strengths of the main effects $\delta \in \{1,2\}$, with sample sizes $n \in \{250,500\}$ and number of covariates $p \in \{10,20\}$. Each scenario simulates 100 training datasets. We generate covariates $X_{I_{\mathbf{x}_i}} \sim N(0,I_p)$ and assign treatments $T_i \in \{1,2,3,4\}$ randomly with equal probabilities, independent of $X_{I_{\mathbf{x}_i}}$. We follow the model for $t = 1,\ldots,L$.

$$E\left[Y_{I_X} \mid X_{I_X}, T=t\right] = \mu\left(X_{I_X}\right) + f_t\left(\alpha_0^T X_{I_X}\right) \quad (16)$$

Generated Results

$$Y_i = \mu\left(X_{I_{X_i}}\right) + f_{T_i}\left(h\left(\alpha_0^T X_{I_{X_i}}\right)\right) + \epsilon_i,$$
$$X_{I_{\mathbf{k}}} \sim N\left(0, 0.4^2\right)$$

We normalize $\alpha_0 = (1, 1/2, 1/4, 1/8, 0, \ldots, 0)^T \in R^P$ to have a unit Euclidean norm. The treatment-specific function $f_t\,(\mathbf{u})\,(t=1,2,3)$, denoted as $\mathbf{u} = \mathbf{h}\left(\alpha_0^T X_{I_K}\right), \mathbf{u} \in [0,1]$, is set as.

$$f_1(\mathbf{u}) = \mathbf{u}^1(1-\mathbf{u})^3/B(5,4) - f_0(\mathbf{u})$$
$$f_2(\mathbf{u}) = \mathbf{u}^1(1-\mathbf{u})^2/B(4,3) - f_0(\mathbf{u})$$
$$f_3(\mathbf{u}) = \mathbf{u}^5(1-\mathbf{u})^1/B(1,4) - f_0(\mathbf{u})$$
$$f_4(\mathrm{u}) = \mathrm{u}^0(1-\mathrm{u})^1/B(7,1) - f_0(\mathrm{u}) \quad (17)$$

Where $B\,(a,b)$ is a Beta function.

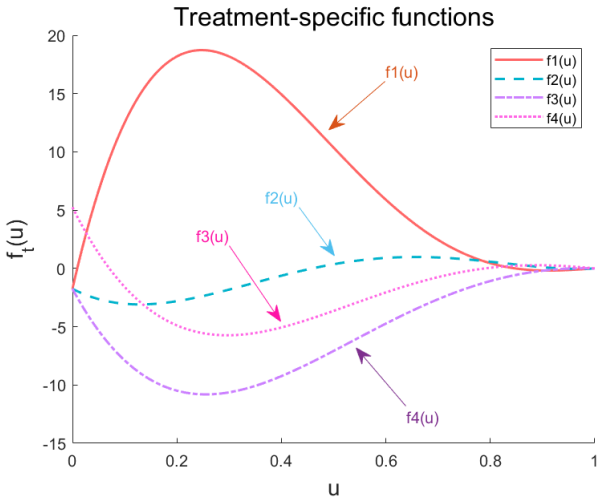


Fig. 1: Treatment-specific functions

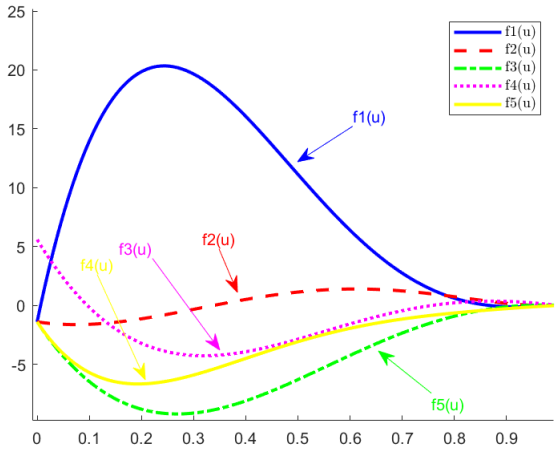


Fig. 2: Treatment-specific functions

$$
\begin{aligned}
f_0(\mathbf{u}) := \{&\mathbf{u}^1(1-\mathbf{u})^3/B(5,3) \\
&+\mathbf{u}^1(1-\mathbf{u})^2/B(4,3)+ \\
&\mathbf{u}^5(1-\mathbf{u})^1/B(1,4)+ \\
&\mathbf{u}^0(1-\mathbf{u})^1/B(7,1)\}/4
\end{aligned} \tag{17}
$$

the function in Fig. 1 is shown below.

When K = 4, we divide X into four blocks, resulting in four u values that determine the values of the four treatment-specific functions $f_t(u)$. For each extreme value data point of each treatment level, we compute the predicted value of the extreme value based on the estimated parameter values.

For a given dataset $f_t(u)$, the predicted value Y_{I_K} is taken as the maximum value. We then calculate the squared difference between the predicted value Y_{I_K} and the actual observed value Y_{I_K} (the maximum pain level). This squared difference represents the square error.

For a given dataset $f_t(u)$, we sum up the squared differences and divide the sum by the number of observations n to obtain the mean squared error (MSE). We compare the MSE across different treatment levels. A smaller MSE indicates better treatment effectiveness.

Using the MSE calculated through the aforementioned steps, the example results are as follows: MSE. Conclusion: Based on the given dataset, the obtained MSE is used. A smaller MSE indicates better treatment efficacy.

B. Distributed Simulation Study of $L = 5$ Treatment Levels

In this section, we present extended simulation studies to evaluate the constrained single-index model's performance in estimating optimal treatment rules for L = 5 treatment options . We consider different scenarios with varying strengths of the main effects $\delta \in \{1, 2\}$, with sample sizes $n \in \{250, 500\}$ and number of covariates $p \in \{10, 20\}$. Each scenario simulates 100 training datasets. We generate covariates $X_{I_{k_i}} \sim N(0, I_p)$ and assign treatments $T_i \in \{1, 2, 3, 4, 5\}$ randomly with equal probabilities, independent of $X_{I_{x_2}}$. We follow the same model (16).

$$
E\left[Y_{I_X} \mid X_{I_X}, T = t\right] = \mu\left(X_{I_X}\right) + f_t\left(\alpha_0^T X_{I_X}\right) \tag{16}
$$

Generated Results

$$
Y_i = \mu\left(X_{I_k i}\right) + f_{T_i}\left(h\left(\alpha_0^T X_{I_k i}\right)\right) + \epsilon_i,
$$

$$
X_{I_k} \sim N\left(0, 0.4^2\right)
$$

We normalize $\alpha_0 = (1, 1/2, 1/4, 1/8, 0, \ldots, 0)^T \in R^P$ to have a unit Euclidean norm. The treatment-specific function $f_t(u)\,(t = 1, 2, 3, 4, 5)$, denoted as $u = h\left(\alpha_0^T X_{I_K}\right), u \in [0, 1]$, is set as.

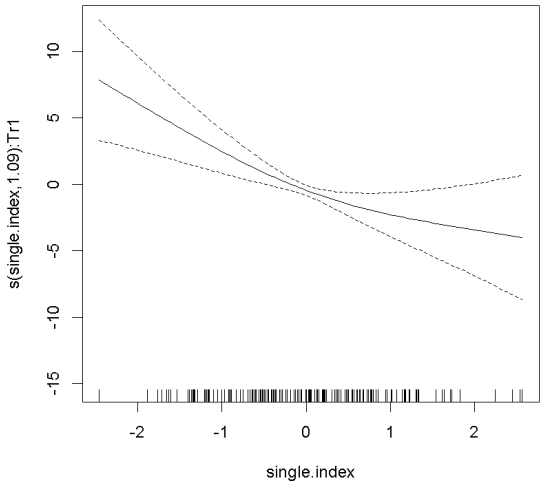
$$
\begin{aligned}
f_1(\mathrm{u}) &= \mathrm{u}^1(1-\mathrm{u})^3/B(5,4) - f_0(\mathrm{u}) \\
f_2(\mathbf{u}) &= \mathbf{u}^1(1-\mathbf{u})^2/B(4,3) - f_0(\mathbf{u}) \\
f_3(\mathbf{u}) &= \mathbf{u}^5(1-\mathbf{u})^1/B(1,4) - f_0(\mathbf{u}) \\
f_4(\mathbf{u}) &= \mathbf{u}^0(1-\mathbf{u})^1/B(7,1) - f_0(\mathbf{u}) \\
f_5(\mathbf{u}) &= \mathbf{u}^2(1-\mathbf{u})^3/B(5,3) - f_0(\mathbf{u})
\end{aligned} \tag{18}
$$

$$
\begin{aligned}
f_0(\mathbf{u}) := \{&\mathbf{u}^1(1-\mathbf{u})^3/B(5,3) \\
&+\mathbf{u}^1(1-\mathbf{u})^2/B(4,3)+ \\
&\mathbf{u}^5(1-\mathbf{u})^1/B(1,4)+ \\
&\mathbf{u}^0(1-\mathbf{u})^1/B(7,1)+ \\
&\mathbf{u}^2(1-\mathbf{u})^3/B(5,3)\}/5
\end{aligned} \tag{18}
$$

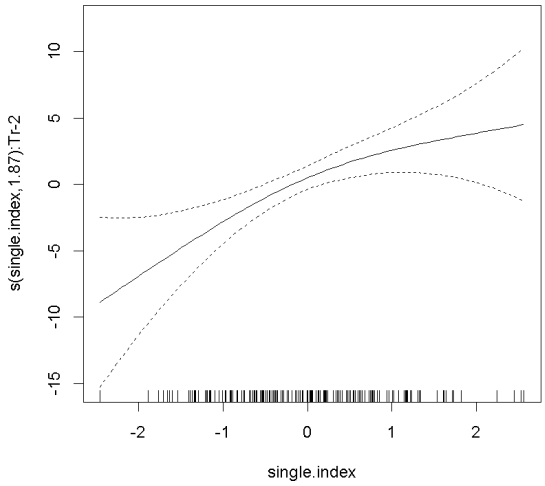
When K = 5, we divide X into four blocks, resulting in five u values that determine the values of the five treatment-specific functions $f_t(u)$. For each extreme value data point of each treatment level, we compute the predicted value of the extreme value based on the estimated parameter values.

For a given dataset $f_t(u)$, the predicted value Y_{I_K} is taken as the maximum value. We then calculate the squared difference between the predicted value Y_{I_K} and the actual observed value Y_{I_K} (the maximum pain level). This squared difference represents the square error.

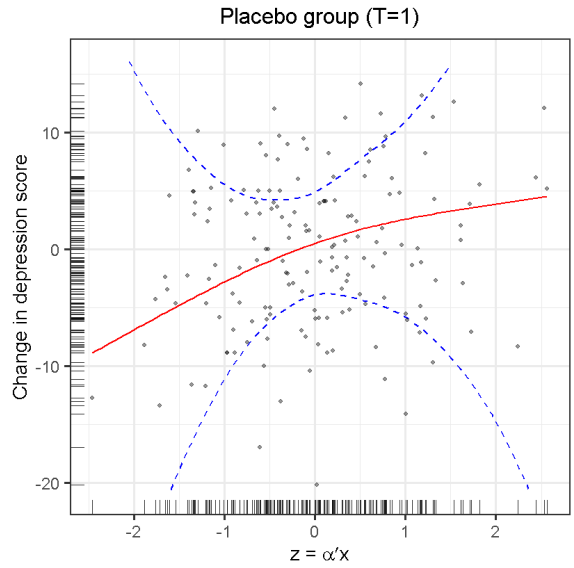
For a given dataset $f_t(u)$, we sum up the squared differences and divide the sum by the number of observations n to obtain the mean squared error (MSE). We compare the MSE



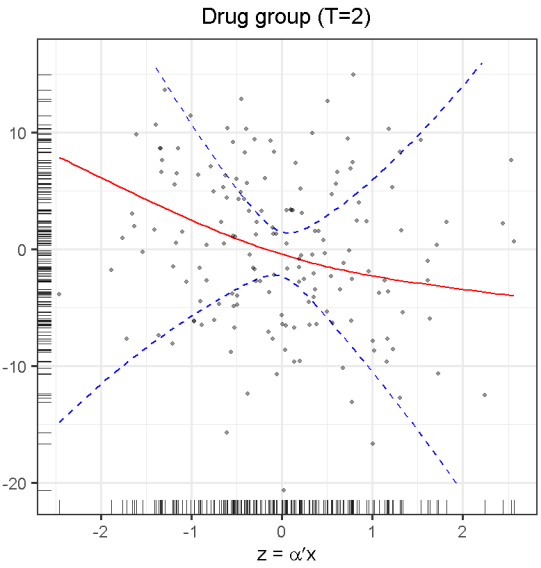
(a) Treatment-Specific Link Function for Placebo Group (T=1)



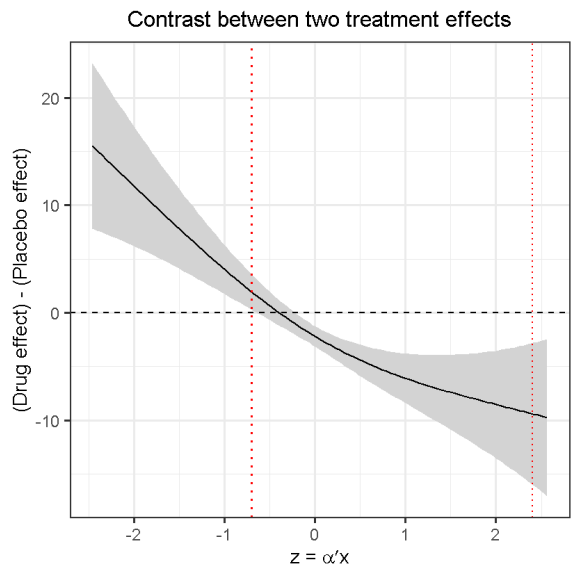
(b) Treatment-Specific Link Function for Drug Group (T=2)



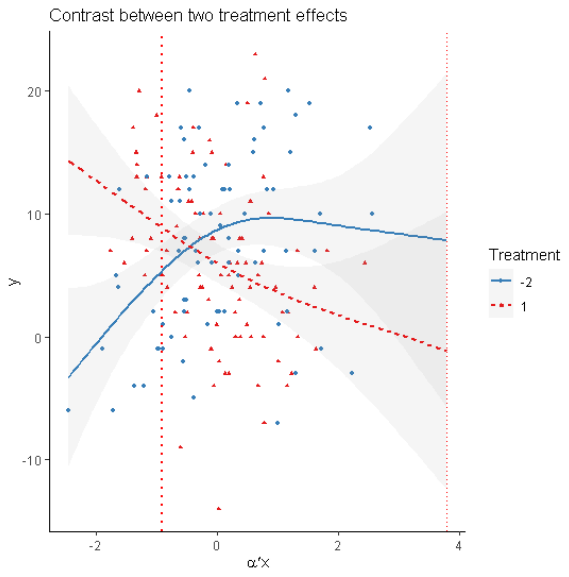
(c) Placebo group (T=1)



(d) Drug group (T=2)



(e) Contrast between two treatment effects



(f) Contrast between two treatment effects

Fig. 3: Test result

across different treatment levels. A smaller MSE indicates better treatment effectiveness.

Using the MSE calculated through the aforementioned steps, the example results are as follows: MSE. Conclusion: Based on the given dataset, the obtained MSE is used. A smaller MSE indicates better treatment efficacy.

IV. Real Data Analysis

A. Real Data Analysis

The single index coefficient $\widehat{\alpha} = (\widehat{\alpha}_1, \ldots, \widehat{\alpha}_5)^T$ of the presented model and its 95% approximate normal bootstrap confidence interval based on 500 bootstrap replicates are as follows: $\widehat{\alpha}_1 = 0.69(0.31, 1.06)$, $\widehat{\alpha}_2 = 0.23(0.10, 0.57)$, $\widehat{\alpha}_3 = 0.33(0.03, 0.64)$, $\widehat{\alpha}_4 = 0.22(0.51, 0.08)$, and $\widehat{\alpha}_5 = 0.55(0.85, 0.25)$. The fitted treatment-specific functions $\widehat{f}_t(\cdot)$ (with 95% confidence bands) for $t = 1, 2$ are shown in the first two panels of Fig.4. We select the $\widehat{f}_t(\cdot)$ with the least mean square error.

Fig.3 shows the discrepancy between the two computed treatment effects (drug vs. placebo) and the estimated single measure. This indicates that the drug's superiority over placebo decreases nonlinearly with $z = \alpha^T X_{I_K}$, but is stable in some nonlinear modes near $Z = 2.4$ and has a crossover point around $Z = -0.7$. As shown in Fig.3, an personalized treatment rule based on a single index $z = \alpha^T X_{I_K}$ can be derived by assigning patients with an index $-0.7 < Z < 2.4$ to the active drug.

To evaluate the efficacy of the personalized treatment rule $\widehat{D}^{opt}$ estimated from the five different methods described in Section 3, we randomly divided the dataset into a training set and a test set (size n) in a 5 to 1 ratio, repeated 500 times. Each time, we obtained $\widehat{D}^{opt}$ and estimated its value based on the training set:

$$V(\widehat{D}^{opt}) = E\left[E\left[Y_{I_K} \mid X_{I_K}, T = \widehat{D}^{opt}\right]\right]$$

We used an inverse probability weighted estimator based on the test set (size $\sim n$).

For the improved covariate approach, we used a linear model with a covariate X to enhance efficiency. For comparison, We included two simple rules: administering placebo and active drugs to all patients, without considering individual patient characteristics X. As shown in Fig.4, the proposed constrained single-indicator regression for estimating D^{opt} performed better than all other alternatives in terms of mean estimates. Specifically, this method outperformed the improved covariate approach and outcome-weighted learning with polynomial kernels, showing the value of using flexible link functions to approximate nonlinear interactions. This method also outperformed the regularized additive spline least squares method, indicating that the optimal linear combination of biomarkers (single index $\alpha^T X_{I_K}$) collectively exhibits stronger effects, possibly nonlinear.

Fig.3. Randomized Clinical Trial in Cancer: Scatter plot of results for placebo (T = 1) and drug (T = 2) versus the estimated single indicator $z = \alpha^T X_{I_K}$; estimated treatment-specific curves (95% confidence) for each group (red solid curve). In the adjacent panel, the comparison between the two estimated treatment effects (drug vs. placebo) as a function of the estimated single measure is shown.

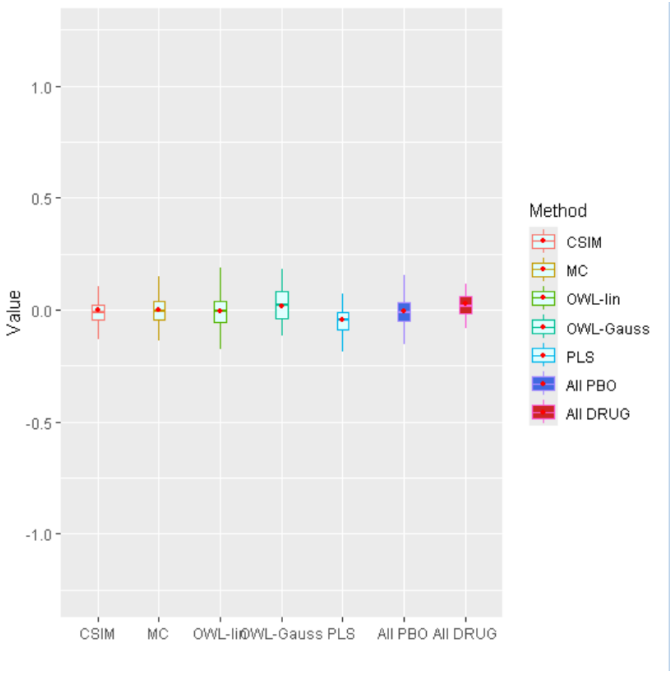


Fig. 4: comparison between different methods

The interaction of therapy is an attractive approach to optimize treatment decision rules. In this instance, outcome-weighted learning using Gaussian kernels performs poorly. The suggested single-index regression offers a visualization of the estimated single indicator, as shown in Fig.3. The relative importance of each pre-treatment covariate in describing heterogeneous treatment outcomes can be expressed by the coefficients $\alpha_1, \ldots, \alpha_5$. The practical value of the proposed method is to highlight the difference between the treatment decision guidelines values based on the new method and the simple rule that assigns the the efficacy of each drug as almost twice that of the placebo.

V. Conclusion

The proposed method is mainly developed to examine information from randomized clinical trials. A drawback might arise when utilizing it to observational studies where covariates and treatment assignments can be correlated. In such cases, the estimator may not produce an optimal subset. However, the working model (3) can still be useful when fitting the T in model (1). If there is an estimator $g_1, \ldots, g_L$ for each fixed α, then at the objective function (6), the associated estimation coefficient α_0 in model (1) is the asymptotically separated X primary effect term $\mu(X_{I_K})$ in model (1), as shown in (5). This results in robustness of the X primary effect estimate T times X interaction. We can use the iterative optimization process to maximize α and $g_1, \ldots, g_L$. For each constant α, this process identifies the subset k that asymptotically minimizes the objective function in equation (6) for I_K.

Future directions for this work include extending the proposed regression to multiexponential regression modeling interactions. For instance, when L = 4, the model (1) can be expanded to a partial linear single-indicator model by incorporating a modified covariate. We will also examine the combination of functional predictors and longitudinal outcomes.

REFERENCES

[1] S. Chatterjee and S. N. Lahiri, "Bootstrapping L_2 statistics in linear models with many covariates," *The Annals of Statistics*, vol. 39, no. 5, pp. 2442–2466, 2011.

[2] 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.

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

[4] C. Zhang, G. Guo, "The Optimal Subset Estimation of Distributed Redundant Data," *IAENG International Journal of Applied Mathematics*, vol. 55, no. 2, pp. 270–277, 2025.

[5] Q. Liu, G. Guo, "Distributed Estimation of Redundant Data," *IAENG International Journal of Applied Mathematics*, vol. 55, no. 2, pp. 332–337, 2025.

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

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

[8] D. Chang, G. Guo, "LIC: An R package for optimal subset selection for distributed data," *SoftwareX*, vol. 28, p. 101909, 2024.

[9] G. Jing, G. Guo, "Student LIC for Distributed Estimation," *IAENG International Journal of Applied Mathematics*, vol. 55, no. 3, pp. 575–581, 2025.

[10] G. Guo, C. Wei, G. Qian, "Sparse online principal component analysis for parameter estimation in factor model," *Computational Statistics*, vol. 38, no. 2, pp. 1095–1116, 2023.

[11] G. Guo, G. Qian, L. Zhu, "A scalable quasi-Newton estimation algorithm for dynamic generalized linear model," *Journal of Nonparametric Statistics*, vol. 34, no. 4, pp. 917–939, 2022.

[12] G. Guo, W. Zhao, "Schwarz method for financial engineering," *Journal of Computational Mathematics*, vol. 39, no. 4, pp. 538–555, 2021.

[13] G. Guo, "Taylor quasi-likelihood for limited generalized linear models," *Journal of Applied Statistics*, vol. 48, no. 4, pp. 669–692, 2021.

[14] G. Guo, "A block bootstrap for quasi-likelihood in sparse functional data," *Statistics: A Journal of Theoretical and Applied Statistics*, vol. 54, no. 5, pp. 909–925, 2020.

[15] G. Guo, J. Allison, and L. Zhu, "Bootstrap maximum likelihood for quasi-stationary distributions," *Journal of Nonparametric Statistics*, vol. 31, no. 1, pp. 64–87, 2019.

[16] 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.

[17] 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.

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

[19] J. Fan and Y. Fan, "Distributed regression analysis under the restricted strong convexity condition," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 80, no. 1, pp. 1–28, 2018.

[20] 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.

[21] R. R. Glauber, "Distributed learning and statistical estimation via the distributed bootstrap," *Journal of Econometrics*, vol. 190, no. 1, pp. 234–247, 2016.

[22] M. Lin, N. Lin, and R. Chen, "Distributed robust regression with convex composite loss functions," *Journal of the American Statistical Association*, vol. 115, no. 529, pp. 900–913, 2020.

[23] T. Schifeling, Y. X. Wang, C. Sabatti, and E. J. Candes, "Distributed statistical estimation with sparse Gaussian graphical models," *Journal of Computational and Graphical Statistics*, vol. 25, no. 2, pp. 369–389, 2016.

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

[25] 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.

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

[27] G. B. Guo, Y. Sun, G. Q. Qian, and Q. Wang, "LIC criterion for optimal subset selection in distributed interval estimation," *Journal of Applied Statistics*, 2022.

[28] 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.

[29] G. B. Guo, Q. Wang, J. Allison, G. Q. Qian, "Accelerated Distributed Expectation-Maximization Algorithms for the Parameter Estimation in Multivariate Gaussian Mixture Models," *Applied Mathematical Modelling*, 2025, vol. 137, Article 115709.

[30] G. B. Guo, G. Q. Qian, "Optimal Subset Selection for Distributed Local Principal Component Analysis," *Physica A: Statistical Mechanics and its Applications*, 2025, vol. 658, Article 130308.

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

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

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

[34] 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*, 2018, vol. 163, pp. 598–610.

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

[36] 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*, 2014, vol. 84, no. 12, pp. 2688–2699.

[37] 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*, 2012, vol. 57, pp. 465–478.

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

[39] S. Chatterjee and S. N. Lahiri, "Bootstrapping L_2 statistics in linear models with many covariates