WILCOXON-TYPE MULTIVARIATE CLUSTER ELASTIC NET

A PREPRINT

Mayu Hiraishi
Graduate School of Culture and Information Science
Doshisha University
Kyoto, Japan
cyag1001@mail4.doshisha.co.jp

Kensuke Tanioka
Department of Biomedical Sciences and Informatics
Doshisha University
Kyoto, Japan
ktanioka@mail.doshisha.ac.jp

Hiroshi Yadohisa
Department of Culture and Information Science
Doshisha University
Kyoto, Japan
hyadohis@mail.doshisha.ac.jp

September 28, 2022

ABSTRACT

We propose a method for high dimensional multivariate regression that is robust to random error distributions that are heavy-tailed or contain outliers, while preserving estimation accuracy in normal random error distributions. We extend the Wilcoxon-type regression to a multivariate regression model as a tuning-free approach to robustness. Furthermore, the proposed method regularizes the $L_1$ and $L_2$ terms of the clustering based on $k$-means, which is extended from the multivariate cluster elastic net. The estimation of the regression coefficient and variable selection are produced simultaneously. Moreover, considering the relationship among the correlation of response variables through the clustering is expected to improve the estimation performance. Numerical simulation demonstrates that our proposed method overperformed the multivariate cluster method and other methods of multiple regression in the case of heavy-tailed error distribution and outliers. It also showed stability in normal error distribution. Finally, we confirm the efficacy of our proposed method using a data example for the gene associated with breast cancer.

Keywords: MM algorithm · multivariate regression · robust statistics

1 Introduction

Numerous studies on multivariate regression analysis have proposed methods to adequately demonstrate the efficiency of the estimation. As part of this framework, it is expected that the accuracy of the estimation is enhanced by considering correlations among the objective variables. Breiman and Friedman [1997] proposed the Curds and Whey method, which predicts multivariate response variables with an optimal linear combination of least squared predictors. Rothman et al. [2010] proposed a multivariate regression with covariance estimation, which estimates regression coefficients and inverse error covariance simultaneously. To improve estimation, multivariate regression has been extended to include a sparse regularization term for high-dimensional data, such as genetic data [Peng et al., 2010, Kim and Xing, 2012, Chen et al., 2017]. These methods consider the correlations of response variables based on a priori data. Methods for reducing the dimension of the response have also been introduced by Cook et al. [2010], Cook and Zhang [2015], and Sun et al. [2015]. The multivariate cluster elastic net, in contrast, is proposed by Price and Sherwood [2018] as a multivariate regression that can take into account the correlation of the response variables without a priori information.

*Use footnote for providing further information about author (webpage, alternative address)—not for acknowledging funding agencies.
This method includes an $L_1$ penalty [Tibshirani, 1996] for the variable selection and $L_2$ penalty for identifying clusters based on $k$-means [Forgy, 1965] that are associated with the response variables. This is different from the previously mentioned methods in that the estimation of the regression coefficients and grouping of the response variables is conducted simultaneously. For data with a small number of samples and a large number of explanatory variables, such as genetic data, the lasso regularization term helps to estimate a sparse coefficient matrix. The multivariate cluster elastic net can be also applied to high dimensional data with multiple response variables. Another characteristic of the multivariate cluster elastic net is the grouping of the fitted values of the response variables by the $k$-means clustering term. With this term, the multivariate cluster elastic net can improve estimation accuracy by considering correlations between the response variables.

Meanwhile, in a regression model, the error distributions are assumed to be Gaussian. However, when the errors follow a heavy-tailed distribution or contain outliers, this may affect the estimation results [e.g., Huber, 2011]. Various robust regression methods have been proposed to solve this problem ([Fan et al., 2017] [Loh, 2017] [Sun et al., 2019] [Lozano et al., 2016] [Wang et al., 2012a] [Avella-Medina and Ronchetti, 2018] [Prasad et al., 2020] [Bellioli et al., 2011] [Bradic et al. 2011] [Wang et al., 2012b] [Wang, 2013] [Fan et al., 2014] [Sun and Zhang, 2012]). Huber regression reduces the impact of outliers by placing an adjustment parameter [Huber, 1964]. Instead of least squares, the lad-lasso ([Wang et al., 2007]) sets the least absolute deviation as its loss function, and Zou and Yuan [2008] introduced the composite quantile regression. The rank-based estimate with Wilcoxon scores [Jureckova, 1971] [Jaeckel, 1972] [Hettmansperger and McKean, 1977] is also known to be more efficient than least squared procedures when the data deviates from a normal distribution. This is because it is robust against a response containing outliers [Hettmansperger and McKean, 1998]. This Wilcoxon estimate does not need a tuning parameter for robust estimation. Wang et al. [2020] and Wang and Li [2009] use a Wilcoxon estimate as their loss function with a regularization penalty for robust estimation in a high-dimensional setting. However, these methods remain in the framework of multiple regression.

We propose a Wilcoxon-type regression that is extended to a multivariate cluster elastic net. It is a high-dimensional regression with robustness to heavy-tailed error distribution or outliers containing random errors. We call this new method the Wilcoxon-type multivariate cluster elastic net (WMCEN). The Wilcoxon-type loss function is expected to be robust to heavy-tailed error distribution or outliers. The proposed method is also suitable for a high-dimensional situation because the variable selection uses the $L_1$ regularization penalty. Moreover, the proposed method obtains more efficiency because the correlation of the response variables is considered, following a similar framework to that of the multivariate cluster elastic net. When updating the parameters of the proposed method, we derive the majorizing function of the updated formula based on the majorize-minimization (MM) algorithm [Hunter and Lange, 2004]. With the majorizing function, an updated formula based on the squared error criterion can be calculated, which ensures the updated formula can be solved easily.

In Section 2 of this paper, we show the objective function of our proposed method, and we explain the derivation of the majorizing function. After describing the algorithm and the updated formula of the proposed method, we show the efficiency of the proposed method through numerical simulation in Section 3. In Section 4, we apply the proposed method to real genetic data related to breast cancer. After discussing based on the results, Section 5 concludes the article with a discussion of further future issues.

## 2 Wilcoxon-type Multivariate Cluster Elastic Net (WMCEN)

In this section, we explain the optimization problem of our proposed method and introduce the majorizing function corresponding to the new method. Before that, we first explain the objective function of Wilcoxon-type regression, which is the basis of the extension of our proposed method. Let $y_i$ $(i = 1, 2, \cdots, n)$ be response variables, where $n$ is the number of subjects. Let $x_i \in \mathbb{R}^p$ $(i = 1, 2, \cdots, n)$ be covariates in $p$ dimension, and $\beta \in \mathbb{R}^p$ be unknown regression coefficients in $p$ dimension. The Wilcoxon-type regression method for a single outcome is estimated by minimizing $\beta$:

$$
\sum_{i < j} |e_i - e_j| \quad (1)
$$

where $e_i = y_i - x_i^T \beta, e_j = y_j - x_j^T \beta$ $(i, j = 1, 2, \cdots, n)$. $e_i$ and $e_j$ are expressed in the form of a linear regression model, and Eq. (1) represents the absolute sum of the difference of the residuals between $i$ and $j$. This function in Eq. (1) is equivalent to Jaeckel’s Wilcoxon-type dispersion function [Jaeckel, 1972]$
\sqrt{12} \sum_{i=1}^{n} \left[ \frac{R(y_i - x_i^T \beta)}{n + 1} - \frac{1}{2} \right] (y_i - x_i^T \beta) \quad (2)$

where $R(y_i - x_i^T \beta)$ is the rank of $(y_i - x_i^T \beta)$ $(i = 1, 2, \cdots, n)$ [Hettmansperger and McKean, 1978]. That is to say, Eq. (2) can be regarded as a weighted regression.
2.1 Optimization problem of Wilcoxon-type multivariate cluster elastic net

Before describing the optimization problem of the proposed method, we introduce the multivariate regression model in this subsection. Let $Y = (y_1, y_2, \ldots, y_q)^T \in \mathbb{R}^{n \times q}$ be matrices of responses, $X = (x_1, x_2, \ldots, x_p)^T \in \mathbb{R}^{n \times p}$ be regression covariates, $B = (\beta_1, \beta_2, \ldots, \beta_q)^T \in \mathbb{R}^{p \times q}$ be unknown regression coefficients, and $E = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)^T$ be i.i.d. random errors. The linear regression model is then represented as follows:

$$Y = XB + E.$$  

Here, we explain the optimization problem of the WMCE. To achieve this, we extend the Wilcoxon-type regression method for each response variable. We extend the Wilcoxon-type regression model to a multivariate regression method as follows:

$$L^*(\{\beta_s\}_{s=1}^q) = \sum_{s=1}^q \sum_{i<j} |e_{is} - e_{js}| \to \min$$  

where $e_{is} = y_{is} - x_i^T \beta_s, e_{js} = y_{js} - x_j^T \beta_s, (i, j = 1, 2, \ldots, n; s = 1, 2, \ldots, q)$. Eq.(3) represents the sum of the objective function of the Wilcoxon-type regression method for each response variable. We extend the Wilcoxon-type regression to the multivariate cluster elastic net (MCEN), and formulate the optimization problem of WMCE based on Eq.(4) as follows:

$$L(\{\beta_s\}_{s=1}^q, U, \{v_\ell\}_{\ell=1}^k) = \sum_{s=1}^q \sum_{i<j} |e_{is} - e_{js}| + \lambda \sum_{s=1}^q \beta_s^1 + \frac{\gamma}{2} \sum_{s=1}^q \sum_{\ell=1}^k u_{s\ell} \|X \beta_s - X v_\ell\|_{F}^2 \to \min$$  

where $\lambda (\lambda > 0)$ and $\gamma$ ($\gamma > 0$) are the tuning parameters. $\| \cdot \|_F$ is the Frobenius norm, $\| \cdot \|_1$ is the $L_1$ norm, and $\| \cdot \|_2$ is the $L_2$ norm. The second and third term of Eq. (4) are the same regularization terms as the multivariate cluster elastic net [Price and Sherwood 2018]. $U = (u_{s\ell}) (s = 1, 2, \ldots, q; \ell = 1, 2, \ldots, k)$ is the indicator matrix, which represents the degree of belonging to cluster $s$ on the response variable $s$. When $u_{s\ell} = 1, X \beta_s$ belongs to the $\ell$th cluster, otherwise $u_{s\ell} = 0$. $v_\ell (\ell = 1, 2, \ldots, k)$ is the partial regression coefficient for cluster centroid. The second formula term is the lasso penalty [Tibshirani 1996], and the third term of Eq.(4) is the $k$-means clustering function.

In this study, we derive the updated formula using the MM algorithm [Hunter and Lange 2004]. To solve the problem of applying the MM algorithm to the lasso penalty [Hunter and Li 2005], we substitute the perturbed version of the lasso penalty term for the traditional lasso penalty term [Yu et al. 2015].

$$L^*(\{\beta_s\}_{s=1}^q, U, \{v_\ell\}_{\ell=1}^k) = \sum_{s=1}^q \sum_{i<j} |e_{is} - e_{js}| + \lambda \sum_{s=1}^q \sum_{h=1}^p (|\beta_{sh}| - \epsilon \log \left(1 + \frac{|\beta_{sh}|}{\epsilon}\right))$$  

where $\epsilon$ ($\epsilon > 0$) is the hyper parameter of the lasso penalty. When $\epsilon$ is approaching 0, the second term of Eq. (5) converges to the second term of Eq. (4).

2.2 Deriving the majorization function of WMCE

In this subsection, we derive the majorizing function from Eq. (5), as shown in Lemma[1] and the majorizing function of the perturbed lasso penalty is derived in Lemma[2]. With Lemma[1] and Lemma[2] we provide the majorizing function of the proposed method in Proposition[3].

We first derive the majorizing function of the first term of Eq. (5). The objective function of WMCE in Eq. (5), that is the first term of Eq. (5), can be transformed as follows:

$$\sum_{s=1}^q \sum_{i<j} |e_{is} - e_{js}| = \sum_{s=1}^q \sum_{i<j} |(y_{is} - x_i^T \beta_s) - (y_{js} - x_j^T \beta_s)|$$  

$$= \sum_{s=1}^q \sum_{i<j} |(y_{is} - y_{js}) - (x_i - x_j)^T \beta_s|$$  

$$= \sum_{s=1}^q \sum_{i<j} |g_{is} - r_{ij}^T \beta_s|$$  

(6)
where \( g_{ijs} = y_{is} - y_{js}, \) and \( r_{ij} = x_i - x_j \) (\( i, j = 1, 2, \ldots, n; \ i < j \)). \( g_{ijs} \) represents the difference between the \( i \) and \( j \) of the \( s \)th response variable, and \( r_{ij} \) is the difference between the \( i \) and \( j \) of the covariate vectors. By forming them in Eq. (6), the difference of the residuals can be expressed as a regression.

Here, we explain the majorizing function. \( \theta \) is set as the parameter of the interested real-valued objective function \( f(\theta) \), and \( \theta^{(t)} \) denotes the fixed value of parameter \( \theta \). In the algorithm, for example, \( \theta^{(t)} \) represents the estimated \( \theta \) at the \( t \)th step. \( g(\theta|\theta^{(t)}) \) also denotes the function, and the \( g(\theta|\theta^{(t)}) \) can easily derive the updated formula when given \( \theta^{(t)} \). The function \( g(\theta|\theta^{(t)}) \) is defined as the majorizing function of \( f(\theta) \) at \( \theta^{(t)} \) when the following two conditions are met:

\[
g(\theta|\theta^{(t)}) \geq f(\theta) \text{ for all } \theta, \\
g(\theta^{(t)}|\theta^{(t)}) = f(\theta^{(t)}). \tag{8}
\]

Next, we obtain the first term of the majorizing function from Eq. (6).

**Lemma 1.** Given \( g_{ijs}, r_{ij}, \) and \( \beta^{(t)}_s \), for any \( \beta_s \), the following inequality holds:

\[
\sum_{s=1}^{q} \sum_{i<j} \left| g_{ijs} - r_{ij} \beta_s \right| \leq \sum_{s=1}^{q} \sum_{i<j} \frac{1}{2} \left| g_{ijs} - r_{ij} \beta^{(t)}_s \right|^2 + \sum_{s=1}^{q} \sum_{i<j} \frac{1}{2} \left| g_{ijs} - r_{ij}^{T} \beta^{(t)}_s \right| \tag{7}
\]

where \( \beta^{(t)}_s \) is the \( t \)th fixed \( \beta_s \) in the algorithm. If \( \beta^{(t)}_s = \beta_s \) for all \( s = 1, 2, \ldots, q \), the equality of Eq. (7) holds.

**Proof.** First, the inequality formula used to derive the majorizing function of the objective function is explained. Eq. (8) is called the arithmetic-geometric mean inequality \[Hunter and Lange [2004]\].

For real numbers \( z_1, z_2, \ldots, z_N \geq 0 \),

\[
\sqrt[N]{\prod_{l=1}^{N} z_l} \leq \frac{1}{N} \sum_{l=1}^{N} z_l, \tag{8}
\]

only when \( z_1 = z_2 = \cdots = z_N \), the equality of Eq. (8) holds.

Here, when \( N = 2, z_1 = |g_{ijs} - r_{ij}^{T} \beta_s|, \) and \( z_2 = |g_{ijs} - r_{ij}^{T} \beta^{(t)}_s|, \) it can be described for any \( i, j, \) and \( s \) by using Eq. (8) as follows:

\[
\sqrt{|g_{ijs} - r_{ij}^{T} \beta_s||g_{ijs} - r_{ij}^{T} \beta^{(t)}_s|} \leq \frac{1}{2} (|g_{ijs} - r_{ij}^{T} \beta_s| + |g_{ijs} - r_{ij}^{T} \beta^{(t)}_s|) \\
\iff |g_{ijs} - r_{ij}^{T} \beta_s||g_{ijs} - r_{ij}^{T} \beta^{(t)}_s| \leq \frac{1}{4} (|g_{ijs} - r_{ij}^{T} \beta_s|^2 + |g_{ijs} - r_{ij}^{T} \beta^{(t)}_s|^2) \\
\iff |g_{ijs} - r_{ij}^{T} \beta_s| \leq \frac{1}{2} \sqrt{|g_{ijs} - r_{ij}^{T} \beta^{(t)}_s|^2 + \frac{1}{2} |g_{ijs} - r_{ij}^{T} \beta^{(t)}_s|^2}. \tag{9}
\]

Because Eq. (9) holds for any \( i, j, \) and \( s \), we have the following inequality:

\[
\sum_{s=1}^{q} \sum_{i<j} |g_{ijs} - r_{ij}^{T} \beta_s| \leq \sum_{s=1}^{q} \sum_{i<j} \frac{1}{2} |g_{ijs} - r_{ij}^{T} \beta^{(t)}_s|^2 + \sum_{s=1}^{q} \sum_{i<j} \frac{1}{2} |g_{ijs} - r_{ij}^{T} \beta^{(t)}_s|. \tag{9}
\]

From **Lemma 1**, the first term of right side in Eq. (5) can be derived as a quadratic function. Then, for any \( \beta_s \), this majorizing function can be described:

\[
\sum_{s=1}^{q} \sum_{o=1}^{n(n-1)/2} \frac{1}{2w_{os}} |g_{os} - r_{o}^{T} \beta_s|^2 + w_{os} \tag{10}
\]

where \( w_{os} = \frac{1}{2|g_{os} - r_{o}^{T} \beta_s|^2} \). Let \( \beta^{(t)}_s \) be a fixed \( p \)-dimensional vector, and the equality of Eq. (10) is established when \( \beta_s = \beta^{(t)}_s \). Because Eq. (10) is a weighted multivariate regression, it is possible to obtain predicted values utilizing the conventional squared error criterion.

The second term of the majorizing function is then provided from the second term of Eq. (5).
Lemma 2. Given $\lambda (\lambda > 0)$, $\epsilon (\epsilon > 0)$, and fixed $\beta_{sh}^{(t)}$, for any $\beta_s$, the following inequality holds:

$$
\lambda \sum_{s=1}^q \sum_{h=1}^p \left( |\beta_{sh}| - \epsilon \log \left( 1 + \frac{|\beta_{sh}|}{\epsilon} \right) \right) \leq \lambda \sum_{s=1}^q \sum_{h=1}^p \left( |\beta_{sh}^{(t)}| - \epsilon \log \left( 1 + \frac{|\beta_{sh}^{(t)}|}{\epsilon} \right) + \frac{(\beta_{sh})^2 - (\beta_{sh}^{(t)})^2}{2(|\beta_{sh}^{(t)}| + \epsilon)} \right)
$$

(11)

where the equality of Eq. (11) holds only when $\beta_{sh} = \beta_{sh}^{(t)}$ for all $\beta_{sh}$.

Proof. Eq. (11) can be proved in the same manner as Yu et al. [2015], and it can be extended to the case of multivariate regression. For any $\beta_{sh}$, the two functions are defined as below:

$$
Q_\epsilon(\beta_{sh}) = \left( |\beta_{sh}| - \epsilon \log \left( 1 + \frac{|\beta_{sh}|}{\epsilon} \right) \right),
$$

$$
S_\epsilon(\beta_{sh}|\beta_{sh}^{(t)}) = |\beta_{sh}^{(t)}| - \epsilon \log \left( 1 + \frac{|\beta_{sh}^{(t)}|}{\epsilon} \right) + \frac{(\beta_{sh})^2 - (\beta_{sh}^{(t)})^2}{2(|\beta_{sh}^{(t)}| + \epsilon)}.
$$

Here, we define the following function:

$$
H_\epsilon(\beta_{sh}|\beta_{sh}^{(t)}) = S_\epsilon(\beta_{sh}|\beta_{sh}^{(t)}) - Q_\epsilon(\beta_{sh}).
$$

(12)

We prove $H_\epsilon(\beta_{sh}|\beta_{sh}^{(t)}) \geq 0$. First, when $\beta_{sh} = \beta_{sh}^{(t)}$,

$$
H_\epsilon(\beta_{sh}^{(t)}|\beta_{sh}^{(t)}) = S_\epsilon(\beta_{sh}^{(t)}|\beta_{sh}^{(t)}) - Q_\epsilon(\beta_{sh}^{(t)})
\quad = \left( |\beta_{sh}^{(t)}| - \epsilon \log \left( 1 + \frac{|\beta_{sh}^{(t)}|}{\epsilon} \right) + \frac{(\beta_{sh})^2 - (\beta_{sh}^{(t)})^2}{2(|\beta_{sh}^{(t)}| + \epsilon)} - \left( |\beta_{sh}^{(t)}| - \epsilon \log \left( 1 + \frac{|\beta_{sh}^{(t)}|}{\epsilon} \right) \right) \right)
\quad = |\beta_{sh}^{(t)}| - \epsilon \log \left( 1 + \frac{|\beta_{sh}^{(t)}|}{\epsilon} \right) + \frac{(\beta_{sh})^2 - (\beta_{sh}^{(t)})^2}{2(|\beta_{sh}^{(t)}| + \epsilon)} - |\beta_{sh}^{(t)}| + \epsilon \log \left( 1 + \frac{|\beta_{sh}^{(t)}|}{\epsilon} \right)
\quad = 0.
$$

(13)

Eq. (13) holds for any $\beta_{sh} = \beta_{sh}^{(t)}$, therefore, the equality of Eq. (11) is satisfied.

Next, when $\beta_{sh} \neq \beta_{sh}^{(t)}$, we prove $H_\epsilon(\beta_{sh}|\beta_{sh}^{(t)}) > 0$ in the cases below:

Case 1 If $0 \leq \beta_{sh} < \beta_{sh}^{(t)}$, $H_\epsilon(\beta_{sh}|\beta_{sh}^{(t)})$ monotonically decreases.

Eq. (12) is differentiated with respect to $\beta_{sh}$,

$$
\frac{\partial H_\epsilon(\beta_{sh}|\beta_{sh}^{(t)})}{\partial \beta_{sh}} = 2 \times \frac{1}{2(|\beta_{sh}^{(t)}| + \epsilon)} \times \beta_{sh} - 1 + \frac{\epsilon}{\epsilon + |\beta_{sh}|}
\quad = \frac{\beta_{sh}}{|\beta_{sh}^{(t)}| + \epsilon} + \frac{\epsilon}{|\beta_{sh}| + \epsilon} - 1.
$$

(14)

For Eq. (14), because $\beta_{sh} < \beta_{sh}^{(t)}$, the following inequality holds:

$$
\frac{\beta_{sh}}{|\beta_{sh}^{(t)}| + \epsilon} + \frac{\epsilon}{|\beta_{sh}| + \epsilon} - 1 < \frac{\beta_{sh}}{|\beta_{sh}| + \epsilon} + \frac{\epsilon}{|\beta_{sh}| + \epsilon} - 1 = \frac{\beta_{sh} + \epsilon}{\beta_{sh} + \epsilon} - 1 = 0.
$$

Therefore,

$$
\frac{\partial H_\epsilon(\beta_{sh}|\beta_{sh}^{(t)})}{\partial \beta_{sh}} = \frac{\beta_{sh}}{|\beta_{sh}^{(t)}| + \epsilon} + \frac{\epsilon}{|\beta_{sh}| + \epsilon} - 1 < 0.
$$

(15)

Eq. (15) shows that $H_\epsilon(\beta_{sh}|\beta_{sh}^{(t)})$ monotonically decreases for $0 \leq \beta_{sh} < \beta_{sh}^{(t)}$. 

5
Case 2 If $0 \leq \beta_{sh}^{(t)} < \beta_{sh}$, $H_e(\beta_{sh} | \beta_{sh}^{(t)})$ monotonically increases. Because $\beta_{sh}^{(t)} < \beta_{sh}$, the following inequality is satisfied:

$$\frac{\partial H_e(\beta_{sh} | \beta_{sh}^{(t)})}{\partial \beta_{sh}} = \frac{\beta_{sh}}{|\beta_{sh}^{(t)}| + \epsilon} + \frac{\epsilon}{|\beta_{sh}^{(t)}| + \epsilon} - 1 > \frac{\beta_{sh}}{|\beta_{sh}| + \epsilon} + \frac{\epsilon}{|\beta_{sh}| + \epsilon} - 1 = \frac{\beta_{sh}}{|\beta_{sh}| + \epsilon} + 1 = 0.$$

Therefore,

$$\frac{\partial H_e(\beta_{sh} | \beta_{sh}^{(t)})}{\partial \beta_{sh}} = \frac{\beta_{sh}}{|\beta_{sh}^{(t)}| + \epsilon} + \frac{\epsilon}{|\beta_{sh}^{(t)}| + \epsilon} - 1 > 0. \quad (16)$$

Eq. (16) shows that $H_e(\beta_{sh} | \beta_{sh}^{(t)})$ monotonically increases when $0 \leq \beta_{sh}^{(t)} < \beta_{sh}$.

Case 3 If $\beta_{sh}^{(t)} < \beta_{sh} \leq 0$, $H_e(\beta_{sh} | \beta_{sh}^{(t)})$ monotonically increases. Eq. (12) is differentiated with respect to $\beta_{sh}$.

$$\frac{\partial H_e(\beta_{sh} | \beta_{sh}^{(t)})}{\partial \beta_{sh}} = 2 \times \frac{1}{2(|\beta_{sh}^{(t)}| + \epsilon)} \times \beta_{sh} - (1) - \frac{\epsilon}{\epsilon + |\beta_{sh}|}$$

$$= \frac{\beta_{sh}}{|\beta_{sh}^{(t)}| + \epsilon} - \frac{\epsilon}{|\beta_{sh}| + \epsilon} + 1.$$

For $\beta_{sh}^{(t)} < \beta_{sh}$, the following inequality holds:

$$\frac{\partial H_e(\beta_{sh} | \beta_{sh}^{(t)})}{\partial \beta_{sh}} = \frac{\beta_{sh}}{|\beta_{sh}^{(t)}| + \epsilon} - \frac{\epsilon}{|\beta_{sh}^{(t)}| + \epsilon} + 1 > \frac{\beta_{sh}}{|\beta_{sh}| + \epsilon} - \frac{\epsilon}{|\beta_{sh}| + \epsilon} + 1.$$

Because $\beta_{sh} \leq 0$,

$$\frac{-|\beta_{sh}| - \epsilon}{|\beta_{sh}| + \epsilon} + 1 = \frac{-(|\beta_{sh}| + \epsilon)}{|\beta_{sh}| + \epsilon} + 1 = 0.$$

Therefore,

$$\frac{\partial H_e(\beta_{sh} | \beta_{sh}^{(t)})}{\partial \beta_{sh}} = \frac{\beta_{sh}}{|\beta_{sh}^{(t)}| + \epsilon} - \frac{\epsilon}{|\beta_{sh}| + \epsilon} + 1 > 0. \quad (17)$$

Eq. (17) indicates that $H_e(\beta_{sh} | \beta_{sh}^{(t)})$ monotonically increases for $0 > \beta_{sh} > -\beta_{sh}^{(t)}$.

Case 4 If $\beta_{sh}^{(t)} < \beta_{sh} \leq 0$, $H_e(\beta_{sh} | \beta_{sh}^{(t)})$ monotonically decreases. Because $\beta_{sh} < -\beta_{sh}^{(t)}$, the following inequality holds:

$$\frac{\partial H_e(\beta_{sh} | \beta_{sh}^{(t)})}{\partial \beta_{sh}} = \frac{\beta_{sh}}{|\beta_{sh}^{(t)}| + \epsilon} - \frac{\epsilon}{|\beta_{sh}^{(t)}| + \epsilon} + 1 < \frac{\beta_{sh}}{|\beta_{sh}| + \epsilon} - \frac{\epsilon}{|\beta_{sh}| + \epsilon} + 1.$$

Because $\beta_{sh} \leq 0$,

$$\frac{-|\beta_{sh}| - \epsilon}{|\beta_{sh}| + \epsilon} + 1 = \frac{-(|\beta_{sh}| + \epsilon)}{|\beta_{sh}| + \epsilon} + 1 = 0.$$

Therefore,

$$\frac{\partial H_e(\beta_{sh} | \beta_{sh}^{(t)})}{\partial \beta_{sh}} = \frac{\beta_{sh}}{|\beta_{sh}^{(t)}| + \epsilon} - \frac{\epsilon}{|\beta_{sh}| + \epsilon} + 1 < 0. \quad (18)$$

Eq. (18) shows that $H_e(\beta_{sh} | \beta_{sh}^{(t)})$ monotonically increases for $\beta_{sh} < \beta_{sh}^{(t)} \leq 0$. 


From Eq. (13) and Cases 1, 2, 3, and 4, for any \( \beta_{sh} \), \( H_x(\beta_{sh}|\beta_{(t)}_{sh}) \geq 0 \). Therefore, Eq. (11) is satisfied.

With Lemma 2, we show the majorizing function can be derived from the perturbed version of the lasso penalty as the squared form. The right-hand side of Eq. (11) can be expressed as

\[
\lambda \sum_{s=1}^{q} \sum_{h=1}^{p} \left| \beta_{sh}^{(t)} \right| - \gamma \log \left( 1 + \frac{\left| \beta_{sh}^{(t)} \right|}{\epsilon} \right) + \frac{\left( \beta_{sh}^{(t)} \right)^2 - \left( \beta_{sh}^{(t)} \right)^2}{2(\left| \beta_{sh}^{(t)} \right| + \epsilon)} = \| \Psi_s \beta_s \|^2_F + C
\]

where

\[
\Psi_s = \text{diag} \left( \frac{1}{\sqrt{2(\beta_{s1}^{(t)} + \epsilon)}}, \frac{1}{\sqrt{2(\beta_{s2}^{(t)} + \epsilon)}}, \ldots, \frac{1}{\sqrt{2(\beta_{sp}^{(t)} + \epsilon)}} \right),
\]

and \( C \) represents constant values that are not relevant to \( \beta_{sh} \). From Lemma 1 and Lemma 2, the following proposition holds.

**Proposition 1.** Given \( Y, X, \lambda (\lambda > 0), \gamma (\gamma > 0) \), and \( \epsilon (\epsilon > 0) \), the following inequality is satisfied:

\[
L^\prime(\{\beta_s\}_{s=1}^{q}, U, \{v_t\}_{t=1}^{k}) = \sum_{s=1}^{q} \sum_{i<j} e_{is} - e_{js} + \lambda \sum_{s=1}^{q} \sum_{h=1}^{p} \left| \beta_{sh} \right| - \gamma \sum_{s=1}^{q} \left( 1 + \left| \beta_{sh} \right| \right) + \frac{\gamma}{2} \sum_{s=1}^{q} \sum_{t=1}^{k} u_{st} \| X \beta_s - X v_t \|_F^2
\]

\[
\leq \sum_{s=1}^{q} \sum_{o=1}^{n(n-1)/2} u_{os} g_{os} - r_o^T \beta_s^2 + \lambda \sum_{s=1}^{q} \left\| \Psi_s \beta_s \right\|_F^2 + C + \frac{\gamma}{2} \sum_{s=1}^{q} \sum_{t=1}^{k} u_{st} \| X \beta_s - X v_t \|_F^2
\]

\[
= M(\{\beta_s\}_{s=1}^{q}, U, \{v_t\}_{t=1}^{k}) \| \beta_s^{(t)} \|_{s=1}^{q}
\]

where the equality holds if \( \beta_s = \beta_s^{(t)} \) in all \( (s, h) (s = 1, 2, \ldots, q; h = 1, 2, \ldots, p) \).

From Eq. (20), the updated formula of the proposed method can be calculated based on the least squared criteria. Therefore, it can be simply expressed.

**2.3 Algorithm and updated formula**

This subsection explains the algorithm and updated formula of the proposed method. The proposed method updates \( \beta_s \), \( U \), and \( v_t \) based on the alternate least squares criterion [1981] by using the majorizing function of the proposed method.

In Proposition 2, we show the updated formula of \( \beta_s \) based on the majorizing function. We then show the procedure for the estimation of \( U \) and \( v_t \) based on k-means [1965]. After that, we present the algorithm of the proposed method.

**Proposition 2.** Given \( X, w_{os}, g_{os}, r_o \) \((o = 1, 2, \ldots, n(n-1)/2; s = 1, 2, \ldots, q), \lambda (\lambda > 0), \gamma (\gamma > 0), \Psi_s, \) and \( u_{st} \) \((s = 1, 2, \ldots, q; t = 1, 2, \ldots, k)\), the updated formula of \( \beta_d \) \((d = 1, 2, \ldots, q)\) is derived as follows:

\[
\beta_{(t+1)}^d \leftarrow \left( 2 \left( \sum_{o=1}^{n(n-1)/2} u_{os}^{(t+1)} r_o^T \right) + 2 \lambda \Psi_d^{(t+1)T} \right) \Psi_d^{(t+1)} + \gamma \sum_{t=1}^{k} u_{dt}^{(t)} X^T X - \gamma \sum_{t=1}^{k} \left( \frac{1}{q} u_{dt}^{(t)} X^T X \right)
\]

where \( q \) indicates the number of \( X \beta_s \) belonging to cluster \( \ell \). \( \beta_s \) is assumed to be updated in the order \( s = 1, 2, \ldots, d, \ldots, q \) as Algorithm 1. \( \beta_{d}^{(t)} \) and \( \beta_{d}^{(t+1)} \) are coefficients at the \( t \)th and \( (t+1) \)th steps, respectively. The function \( f(m, s) \) is defined as \( t + 1 \) when \( m < s \), whereas it is defined as \( t \) when \( m > s \).
\textbf{Proof.} We provide the results of the derivatives of each term on the right-hand side of Eq. (20) and an updated formula based on these results.

For the first term on the right side of Eq. (20),

\[ M_1(\beta_s) = \sum_{s=1}^{q} \frac{n(n-1)/2}{2} w_{os} |g_{os} - r_o^T \beta_s|^2 \]

\[ = \sum_{s=1}^{q} \frac{n(n-1)/2}{2} w_{os} (g_{os}^2 - 2g_{os} r_o^T \beta_s + \beta_s^T r_o r_o^T \beta_s) \]

\[ = \sum_{s=1}^{q} \frac{n(n-1)/2}{2} w_{os} g_{os}^2 - 2 \sum_{s=1}^{q} \frac{n(n-1)/2}{2} w_{os} g_{os} r_o^T \beta_s + \sum_{s=1}^{q} \frac{n(n-1)/2}{2} w_{os} r_o r_o^T \beta_s \]

where the second and third terms in Eq. (22) are related to \( \beta_s \).

Here, Eq. (22) is differentiated by \( \beta_d \) \((d = 1, 2, \ldots, q)\), and we have

\[ \frac{\partial M_1(\beta_s)}{\partial \beta_d} = -2 \left( \sum_{o=1}^{n(n-1)/2} w_{od} g_{od} r_o \right) + 2 \left( \sum_{o=1}^{n(n-1)/2} w_{od} r_o r_o^T \right) \beta_d. \]  

Next, the second term of Eq. (20) can be described as follows:

\[ M_2(\beta_s) = \lambda \sum_{s=1}^{q} \| \Psi_s \beta_s \|_F^2 \]

\[ = \lambda \sum_{s=1}^{q} \left( \beta_s^T \Psi_s^T \Psi_s \beta_s \right) \]

\[ = \sum_{s=1}^{q} \beta_s^T (\lambda \Psi_s^T \Psi_s) \beta_s \]

Here, Eq. (24) is differentiated by \( \beta_d \), and we have

\[ \frac{\partial M_2(\beta_s)}{\partial \beta_d} = 2 \lambda \Psi_s^T \Psi_d \beta_d. \]

Then, the third term of the right side of Eq. (20) can be described as follows:

\[ M_3(\beta_s) = \frac{\gamma}{2} \sum_{s=1}^{q} \sum_{\ell=1}^{k} u_{s,\ell} \| X \beta_s - X v_{\ell} \|^2 \]

\[ = \frac{\gamma}{2} \sum_{s=1}^{q} \sum_{\ell=1}^{k} u_{s,\ell} (X \beta_s - X v_{\ell})^T (X \beta_s - X v_{\ell}) \]

\[ = \frac{\gamma}{2} \sum_{s=1}^{q} \sum_{\ell=1}^{k} u_{s,\ell} (\beta_s^T X^T X \beta_s - 2 \beta_s^T X^T X v_{\ell} + v_{\ell}^T X^T X v_{\ell}) \]

\[ = \frac{\gamma}{2} \sum_{s=1}^{q} \sum_{\ell=1}^{k} u_{s,\ell} (\beta_s^T X^T X \beta_s - \gamma \sum_{s=1}^{q} \sum_{\ell=1}^{k} u_{s,\ell} \beta_s^T X^T X v_{\ell} + \frac{\gamma}{2} \sum_{s=1}^{q} \sum_{\ell=1}^{k} u_{s,\ell} v_{\ell}^T X^T X v_{\ell}) \]

Before deriving the updated formula, to find \( v_{\ell} \) for the optimization of the right side of Eq. (20), we have:

\[ \frac{\partial M_3(\beta_s)}{\partial v_{\ell}} = -\gamma \sum_{s=1}^{q} u_{s,\ell} X^T X \beta_s + \gamma \sum_{s=1}^{q} u_{s,\ell} X^T X v_{\ell} = 0_p \]

\[ \iff -\gamma \sum_{s=1}^{q} u_{s,\ell} X^T X \beta_s + \gamma \sum_{s=1}^{q} u_{s,\ell} X^T X v_{\ell} = 0_p \]

\[ \iff \sum_{s=1}^{q} u_{s,\ell} X^T X v_{\ell} = \sum_{s=1}^{q} u_{s,\ell} X^T X \beta_s \]

\[ \iff v_{\ell} = \frac{1}{q_{\ell}} \sum_{s=1}^{q} u_{s,\ell} \beta_s. \]  

8
where $O_p$ is the element is 0 with the length of $p$.

Here, when substituting Eq. (27) into Eq. (28), we have

\[
M_3^q(\{\beta_s\}_{s=1}^q) = \frac{1}{2} \sum_{\ell=1}^k \sum_{s=1}^q u_{s\ell} \beta_s^T X^T X \beta_s - \gamma \sum_{\ell=1}^k \sum_{s=1}^q u_{s\ell} \beta_s^T X^T X \left( \frac{1}{q_{\ell}} \sum_{m=1}^q u_{m\ell} \beta_m \right)
+ \gamma \frac{1}{2} \sum_{\ell=1}^k \sum_{s=1}^q u_{s\ell} \left( \frac{1}{q_{\ell}} \sum_{m=1}^q u_{m\ell} \beta_m \right)^T X^T X \left( \frac{1}{q_{\ell}} \sum_{m=1}^q u_{m\ell} \beta_m \right) \tag{28}
\]

To derive the updated formula of $\beta_d$, the second term of Eq. (28) can be divided into a term relevant to $\beta_d$ and a term not relevant to $\beta_d$:
Similarly, the third term of Eq. (28) can be divided into a term relevant to $\beta_d$ and one not relevant to $\beta_d$.

\[
M_3^i\{\{\beta_s\}_{s=1}^q\} = \frac{\gamma}{2} \sum_{q\ell} \sum_{m=1}^q \left( \frac{1}{q\ell} \sum_{m,d} u_{m\ell} \beta_m \right) X^T X \left( \frac{1}{q\ell} \sum_{m=1}^q u_{m\ell} \beta_m \right)
\]

\[
= \frac{\gamma}{2} \sum_{q\ell} \left( \frac{1}{q\ell} \sum_{m,d} u_{m\ell} \beta_d + \frac{1}{q\ell} \sum_{m \neq d} u_{m\ell} \beta_m \right) X^T X \left( \frac{1}{q\ell} \sum_{m,d} u_{m\ell} \beta_d + \frac{1}{q\ell} \sum_{m \neq d} u_{m\ell} \beta_m \right)
\]

\[
= \frac{\gamma}{2} \sum_{q\ell} \left( \frac{1}{q\ell} \sum_{m,d} u_{m\ell} \beta_d + \frac{1}{q\ell} \sum_{m \neq d} u_{m\ell} \beta_m \right) X^T X \left( \frac{1}{q\ell} \sum_{m,d} u_{m\ell} \beta_d \right)
\]

\[
\quad + \frac{\gamma}{2} \sum_{q\ell} \left( \frac{1}{q\ell} \sum_{m \neq d} u_{m\ell} \beta_d \right) X^T X \left( \frac{1}{q\ell} \sum_{m,d} u_{m\ell} \beta_d \right)
\]

\[
= \beta_d^T \left( \sum_{q\ell} \frac{1}{q\ell} u_{m\ell} X^T X \beta_d \right) \beta_d + \beta_d^T \left( \sum_{q\ell} \frac{1}{q\ell} u_{m\ell} \beta_m \right)
\]

where the first and second terms of Eq. (30) are relevant to $\beta_d$. Using Eq. (29) and Eq. (30), Eq. (28) can be described as follows:

\[
M_3^i\{\{\beta_s\}_{s=1}^q\} = \frac{\gamma}{2} \sum_{q\ell} \sum_{m=1}^q \left( \frac{1}{q\ell} \sum_{m,d} u_{m\ell} \beta_m \right) X^T X \beta_s - \beta_d^T \left( \sum_{q\ell} \frac{1}{q\ell} u_{m\ell} \beta_m \right) \beta_d - \beta_d^T \left( 2\gamma \sum_{q\ell} \frac{1}{q\ell} u_{m\ell} \beta_m \right)
\]

\[
\quad - \gamma \sum_{q\ell} \sum_{m \neq d} u_{m\ell} \beta^T \left( \sum_{q\ell} \frac{1}{q\ell} u_{m\ell} \beta_m \right) \beta_d + \gamma \sum_{q\ell} \sum_{m \neq d} u_{m\ell} \beta^T \left( \sum_{q\ell} \frac{1}{q\ell} u_{m\ell} \beta_m \right)
\]

\[
\quad + \gamma \sum_{q\ell} \sum_{m \neq d} u_{m\ell} \beta^T \left( \sum_{q\ell} \frac{1}{q\ell} u_{m\ell} \beta_m \right) \beta_d + \gamma \sum_{q\ell} \sum_{m \neq d} u_{m\ell} \beta^T \left( \sum_{q\ell} \frac{1}{q\ell} u_{m\ell} \beta_m \right)
\]

\[
\quad + \gamma \sum_{q\ell} \sum_{m \neq d} u_{m\ell} \beta^T \left( \sum_{q\ell} \frac{1}{q\ell} u_{m\ell} \beta_m \right)
\]

\[
= \frac{\gamma}{2} \sum_{q\ell} u_{m\ell} \beta^T \left( \sum_{q\ell} \frac{1}{q\ell} u_{m\ell} \beta_m \right) X^T X \beta_d - \beta_d^T \left( \sum_{q\ell} \frac{1}{q\ell} u_{m\ell} \beta_m \right) \beta_d - \beta_d^T \left( 2\gamma \sum_{q\ell} \frac{1}{q\ell} u_{m\ell} \beta_m \right)
\]

\[
\quad + \beta_d^T \left( \sum_{q\ell} \frac{1}{q\ell} u_{m\ell} \beta_m \right) \beta_d + \beta_d^T \left( \sum_{q\ell} \frac{1}{q\ell} u_{m\ell} \beta_m \right)
\]

\[
= \beta_d^T \left( \sum_{q\ell} \frac{1}{q\ell} u_{m\ell} X^T X \beta_d \right) - \beta_d^T \left( \sum_{q\ell} \frac{1}{q\ell} u_{m\ell} \beta_m \right)
\]

\[
+ \beta_d^T \left( \sum_{q\ell} \frac{1}{q\ell} u_{m\ell} \beta_m \right) \beta_d
\]

\[
= \beta_d^T \left( \sum_{q\ell} \frac{1}{q\ell} u_{m\ell} X^T X \beta_d \right) - \beta_d^T \left( \sum_{q\ell} \frac{1}{q\ell} u_{m\ell} \beta_m \right)
\]

\[
+ \beta_d^T \left( \sum_{q\ell} \frac{1}{q\ell} u_{m\ell} \beta_m \right) \beta_d
\]

\[
- \beta_d^T \left( \sum_{q\ell} \frac{1}{q\ell} u_{m\ell} \beta_m \right) + C_2
\]

\[
(31)
\]
where $C_2$ is the constant value that is not relevant to $\beta_d$.

Here, Eq. (31) is differentiated by $\beta_d$ ($d = 1, 2, \cdots, q$), and we have

$$
\frac{\partial M^s((\beta_s)_{s=1}^q)}{\partial \beta_d} = \gamma \sum_{\ell=1}^k u_{d\ell} X^T X \beta_d - \gamma \sum_{\ell=1}^k \frac{1}{q_{\ell}} u_{d\ell} X^T X \beta_d - \gamma \sum_{\ell=1}^k u_{d\ell} \frac{1}{q_{\ell}} \sum_{m \neq d} q \ u_{m\ell} X^T X \beta_d
$$

(32)

From the above, with Eq. (23), Eq. (25), and Eq. (32), we have the updated formula of $\beta_d$ as follows:

$$
\frac{\partial M^s((\beta_s)_{s=1}^q)}{\partial \beta_d} = -2 \left( \sum_{o=1}^{(n-1)/2} w_{o\ell} g_{o\ell} r_o^T \right) + 2 \left( \sum_{o=1}^{(n-1)/2} w_{o\ell} \Psi_o r_o^T \right) \beta_d + 2 \left( \lambda \Psi_d^T \Psi_d \right) \beta_d
$$

$$
\begin{align*}
&= 2 \left( \sum_{o=1}^{(n-1)/2} w_{o\ell} g_{o\ell} r_o^T \right) + 2 \left( \lambda \Psi_d^T \Psi_d \right) + \gamma \sum_{\ell=1}^k u_{d\ell} X^T X - \gamma \sum_{\ell=1}^k \frac{1}{q_{\ell}} u_{d\ell} X^T X \beta_m
\end{align*}
$$

$$
\iff \beta_d = 2 \left( \sum_{o=1}^{(n-1)/2} w_{o\ell} g_{o\ell} r_o^T \right) + 2 \left( \lambda \Psi_d^T \Psi_d \right) + \gamma \sum_{\ell=1}^k u_{d\ell} X^T X - \gamma \sum_{\ell=1}^k \frac{1}{q_{\ell}} u_{d\ell} X^T X \beta_m
$$

(33)

As Proposition 2, because $\beta_s$ is updated in the order, function $f(m, s)$ is defined as $t + 1$ when $m < s$, whereas it is defined as $t$ when $m > s$.

In Algorithm 1, when updating $\beta_s$, $u_{s\ell}$ is fixed as $u_{s\ell}^t$. Therefore, the updated formula of $\beta_s$ is obtained as Eq. (21).

Next, we explain the updated rule of $U$ and the updated formula of $v_\ell$ ($\ell = 1, 2, \cdots, k$). $U$ and $v_\ell$ are updated in the same manner as the $k$-means.

**Update $U$**

$U$ is updated for each $s$. $u_{s\ell}$ is updated as

$$
u_{s\ell}^{(t+1)} \leftarrow \begin{cases} 1 & \|X \beta_s^{(t)} - X v_\ell^{(t)}\|_F^2 \leq \|X \beta_s^{(t)} - X v_\ell^{(t)}\|_F^2 \text{ for any } \ell (\ell = 1, 2, \cdots, k) \\ 0 & \text{(otherwise)} \end{cases}
$$

(34)

which is applied to all $s$ ($s = 1, 2, \cdots, q$) and $\ell$ ($\ell = 1, 2, \cdots, k$).

**Update $v_\ell$ ($\ell = 1, 2, \cdots, k$)**

$v_\ell$ is updated as

$$
u_\ell^{(t+1)} \leftarrow \frac{1}{q_{\ell}} \sum_{s=1}^q u_{s\ell}^{(t+1)} \beta_s^{(t)} \quad (\ell = 1, 2, \cdots, k).
$$

(35)

Our proposed method updates $\beta_s$, auxiliary variables $w_{o\ell}$ and $\Psi_s$, indicator matrix $U$, and cluster centroid coefficient $v_\ell$ alternatively based on the MM algorithm. The details of the MM algorithm are described in Algorithm 1.
Algorithm 1 Wilcoxon-type MCEN based on MM algorithm

**Require:** $X, Y, k, \lambda > 0, \gamma > 0, \epsilon > 0$; threshold for this algorithm $\epsilon^* > 0$

**Ensure:** $\beta(s = 1, 2, \cdots, q), U, v_i(\ell = 1, 2, \cdots, k)$

1: Set $t \leftarrow 1$ and $t^\dagger \leftarrow 1$
2: for $s = 1$ to $q$ do
3:     Set initial values $\beta_s^{(t)}$
4: end for
5: Set initial $U^{(t)}$ and $v_i^{(t)}(\ell = 1, 2, \cdots, k)$ by $k$-means
6: while $L^t(\{\beta_s^{(t)}\}_{s=1}^q, U^{(t)}), (v_i^{(t)}(\ell = 1, 2, \cdots, k)) - L^t(\{\beta_s^{(t+1)}\}_{s=1}^q, U^{(t+1)}), (v_i^{(t+1)}(\ell = 1, 2, \cdots, k)) \geq \epsilon^*$ do
7:     while $L^t(\{\beta_s^{(t)}\}_{s=1}^q, U^{(t)}), (v_i^{(t)}(\ell = 1, 2, \cdots, k)) - L^t(\{\beta_s^{(t+1)}\}_{s=1}^q, U^{(t+1)}), (v_i^{(t+1)}(\ell = 1, 2, \cdots, k)) \geq \epsilon^*$ do
8:         for $s = 1$ to $q$ do
9:             for $\alpha = 1$ to $(n(n-1))/2$ do
10:                 $w_{\alpha s}^{(t+1)} \leftarrow 1/(2|g_{\alpha s} - r_{\alpha g}^\dagger \beta_s^{(t)}|^q)$
11:             end for
12:             $\Psi^{(t+1)} \leftarrow \text{diag} \left( \frac{1}{\sqrt{2(\beta_{s1}^{(t)}+\epsilon)}}, \frac{1}{\sqrt{2(\beta_{s2}^{(t)}+\epsilon)}}, \cdots, \frac{1}{\sqrt{2(\beta_{sk}^{(t)}+\epsilon)}} \right)$
13:             $\beta_s^{(t+1)} \leftarrow \left(2 \sum_{\alpha=1}^{n(n-1)/2} \frac{\Psi^{(t+1)} \Psi^{(t+1)}}{w_{\alpha s}^{(t+1)} g_{\alpha s} r_o} + 2 \sum_{\alpha=1}^{n(n-1)/2} \gamma \sum_{t=1}^k \frac{1}{q^t} v_i^{(t)} X^T X - \gamma \sum_{t=1}^k \frac{1}{q^t} u_i^{(t)} X^T X \right)^{-1}$
14:         end for
15:     end while
16: while $L^t(\{\beta_s^{(t)}\}_{s=1}^q, U^{(t)}), (v_i^{(t)}(\ell = 1, 2, \cdots, k)) - L^t(\{\beta_s^{(t+1)}\}_{s=1}^q, U^{(t+1)}), (v_i^{(t+1)}(\ell = 1, 2, \cdots, k)) \geq \epsilon^*$ do
17:         for $s = 1$ to $q$ do
18:             Update $U^{(t)}$ to $U^{(t+1)}$ based on Eq. (34)
19:         end for
20: end while
21: Update $v_i^{(t)}$ to $v_i^{(t+1)}$ based on Eq. (35)
22: end while
23: $t^\dagger \leftarrow t^\dagger + 1$
24: end while
25: end while
26: end while

3 Numerical Simulation

In this section, we demonstrate the efficiency of the proposed method through a numerical simulation.

3.1 Simulation design

This simulation generates covariate matrix $X$ and response variables $Y$ with a true coefficient matrix. We show the simulation setting patterns in Table 1. The simulation evaluates the prediction accuracy and mean squared error (MSE) of the coefficient matrix. In this simulation, we use RStudio Version 1.4.1103.

We generate the simulation data, adjusted based on Price and Sherwood [2018]. The sample size is 50, and the number of response variables is 9. We set the number of covariate matrix $p$ as 12 and 100, which will be discussed in Factor 2. First, let

$$\Sigma = (\tilde{\sigma}_{jj}), \tilde{\sigma}_{jj} = 1, \tilde{\sigma}_{j'j} = 0.7, (j, j' = 1, 2, \cdots, 12).$$  \hspace{1cm} (36)

The covariate vectors are generated by $X_i \sim \mathcal{N}(0_{12}, \Sigma)$ for $p = 12$. In the case of $p = 100$, $X_i \sim \mathcal{N}(0_{100}, \Sigma')$, where

$$\Sigma' = \begin{pmatrix} \tilde{\Sigma} & O_{12,88} \\ O_{88,12} & I_{88} \end{pmatrix}. \hspace{1cm} (37)$$

$O_{12,88}$ is a zero matrix with 12 rows and 88 columns, $O_{88,12}$ is a zero matrix with 88 rows and 12 columns, and $I_{88}$ is the identity matrix.
To make the matrix of $B$, in the case of $p = 12$, we set the $b_4(\eta, \xi) = (\eta_4 - \xi, \eta_4, \eta_4 + \xi) \in \mathbb{R}^{4 \times 3}$. $\eta_k$ is a vector with the length of $K$, whose elements are all $\eta$. Details on how the values of $\eta$ and $\xi$ are determined will be explained in Factor 3 and Factor 4, respectively.

Here, the true $B^{*}_{\eta, \xi}$ in $p = 12$ is set as follows:

$$B^{*}_{\eta, \xi} = \begin{pmatrix} b_4(\eta, \xi) & O_{4,3} & O_{4,3} \\ O_{4,3} & b_4(\eta, \xi) & O_{4,3} \\ O_{4,3} & O_{4,3} & b_4(\eta, \xi) \end{pmatrix}. \quad (38)$$

For the coefficient groups in $p = 100$, as is the case of $p = 12$, $b_{10}(\eta, \xi) = (\eta_{10} - \xi, \eta_{10}, \eta_{10} + \xi) \in \mathbb{R}^{10 \times 3}$.

$$B^{*}_{\eta, \xi} = \begin{pmatrix} b_{10}(\eta, \xi) & O_{10,3} & O_{10,3} \\ O_{10,3} & b_{10}(\eta, \xi) & O_{10,3} \\ O_{10,3} & O_{10,3} & b_{10}(\eta, \xi) \end{pmatrix}. \quad (39)$$

Next, we explain how to generate the response variable, with $X, B$. The response variable is generated as follows:

$$y_i = B^{*}_{\eta, \xi}^T X_i + \varepsilon_i. \quad (40)$$

Here, $\varepsilon_i$ is the random error. Several different settings are employed, which are described in detail in Factor 5. The number of total patterns in this simulation setting is $4 \times 2 \times 4 \times 3 \times 4 \times 4 = 384$. For each pattern combination, we generate 50 learning samples and 1000 test samples from Eq. (40), and we repeat the calculation 100 times.

Next, we explain the evaluation index. In this simulation, we evaluate the prediction variable and coefficient matrix. The median of the absolute prediction error (APE) is

$$\text{median}\{|y_{is}^* - \hat{y}_{is}|, \ i = 1, 2, \cdots, 1000; s = 1, 2, \cdots, 9\}$$

where $y_{is}^*$ represents the testing samples and $\hat{y}_{is}$ represents the prediction variables. We also compare the mean squared error of estimator

$$\frac{1}{p \times 9} \sum_{s=1}^9 \| \hat{\beta}_s - \beta_s^* \|_2^2$$

where $\beta_s^*$ is the true $\beta_s$, and $\hat{\beta}_s$ are the predicted coefficient vectors.

Now, we explain each simulation setting factor denoted in Table 1.

**Factor 1: Method**

We apply the 4 methods to compare their performance.

In addition to the proposed method, we apply three methods for comparison: the multivariate cluster elastic net (MCEN) [Price and Sherwood, 2018], separate elastic net [Zou and Hastie, 2005], and Wilcoxon-type lasso [Wang et al., 2020]. The optimization problem of MCEN is shown as follows:

$$\frac{1}{2n_0} \| Y - XB \|_F^2 + \lambda \sum_{s=1}^9 \| \beta_s \|_1 + \frac{\gamma}{2} \sum_{s=1}^9 \sum_{\ell=1}^k u_{s\ell} \| X \beta_s - X v_\ell \|_2^2$$

where $k$ is the number of clusters, and $\lambda (\lambda > 0)$ and $\gamma (\gamma > 0)$ are the tuning parameters. $U = (u_{s\ell}) (s = 1, 2, \cdots, 9; \ell = 1, 2, \cdots, k)$ is the indicator matrix representing the degree of belonging to cluster $\ell$ on the response.

| Factor No. | Factor name                  | level |
|------------|------------------------------|-------|
| Factor 1   | Method                       | 4     |
| Factor 2   | Number of covariate variable | 2     |
| Factor 3   | Number of parameter $\eta$ for true coefficient matrix | 4     |
| Factor 4   | Number of parameter $\xi$ for true coefficient matrix | 3     |
| Factor 5   | Number of error distribution | 4     |
variable \( s \cdot v_\ell (\ell = 1, 2, \cdots, k) \) is the partial regression coefficient for the cluster centroid. Elastic net is applied to each response variable separately, denoted as SEN. The objective function of elastic net is as follows:
\[
\| y_s - X \beta_s \|^2 + \delta_s \| \beta_s \|_1 + \gamma_s \| \beta_s \|^2_2 \quad (s = 1, 2, \cdots, q)
\]
where \( \delta_s (\delta_s > 0) \) and \( \gamma_s (\gamma_s > 0) \) are tuning parameters. The third compared method is lasso regression, for which the loss function is a Wilcoxon-type regression function. It is denoted by WLASSO. The objective function of WLASSO is as follows:
\[
\sum_{i<j} | e_{is} - e_{js} | + \gamma_s \| \beta_s \|_1 \quad (s = 1, 2, \cdots, q)
\]
where \( e_{is} = y_{is} - x_i^T \beta_s, e_{js} = y_{js} - x_j^T \beta_s \), and \( \gamma_s (\gamma_s > 0) \) is the tuning parameter.

MCEN uses the \texttt{mcen} package in R software [Sherwood and Price, 2020], and SEN are fitted by the \texttt{glmnet} package in R software [Friedman et al., 2008]. For WLASSO, we use \texttt{LADlasso} [Wang et al., 2007] in the package called \texttt{MTE} [Li and Qin, 2021], as mentioned in Wang et al. [2020]. The tuning parameters of all methods are chosen by five-fold cross validation, and the cluster sizes of 2 and 3 are candidates for the proposed method and MCEN.

Factor 2: Covariate variable
The number of the covariate vectors is set as \( p = 12 \) and 100. The covariate matrix for \( p = 12 \) is described in Eq. (36), and it is described for \( p = 100 \) in Eq. (37).

Factor 3: Parameter \( \eta \) for true coefficient matrix \( B^*_{\eta, \xi} \)
\( \eta \) is a factor to control each value of the non-zero part of a true coefficient matrix. The candidates for \( \eta \), one of the parameters for the true coefficient matrix, are set as 0.25, 0.5, 0.75, and 1.

Factor 4: Parameter \( \xi \) for true coefficient matrix \( B^*_{\eta, \xi} \)
\( \xi \) is a factor to control the difference between each non-zero true coefficient vector in \( B^*_{\eta, \xi} \). The candidates for \( \xi \), the other parameter for the true coefficient matrix, are set as 0.02, 0.05, and 0.10. With \( \eta \) and \( \xi \), the true coefficient matrix generates \( B^*_{\eta, \xi} \) in Eq. (38) and Eq. (39).

Factor 5: Error distribution
We set four different patterns for \( \varepsilon_i = (\varepsilon_{is}) \) based on Wang et al. [2020].
Error 1 is set as
\[
\varepsilon_{is} \sim N(0, 1) \quad (i = 1, 2, \cdots, n; s = 1, 2, \cdots, q).
\]
Error 2 is set as
\[
\varepsilon_{is} \sim 0.95N(0, 1) + 0.05N(0, 100) \quad (i = 1, 2, \cdots, n; s = 1, 2, \cdots, q).
\]
Error 3 is set as
\[
\varepsilon_{is} \sim \sqrt{2}t(4) \quad (i = 1, 2, \cdots, n; s = 1, 2, \cdots, q)
\]
where \( t(4) \) represents the \( t \) distribution with 4 degree of freedom.
Error 4 is set as
\[
\varepsilon_{is} \sim Cauchy(0, 1) \quad (i = 1, 2, \cdots, n; s = 1, 2, \cdots, q)
\]
where \( Cauchy(0, 1) \) represents a Cauchy distribution with a location parameter of 0 and a scale parameter of 1. Error 1 assumes a normal distribution, Error 2 assumes a distribution containing outliers, and Error 3 and Error 4 assume heavy-tailed distribution.
3.2 Simulation result

The results of the simulation are displayed by error distribution in Tables 2 to 5 for $p = 12$, and Tables 6 to 9 for $p = 100$. First, as seen by each outcome for Factor 1, the results of the proposed method were better than those of the other compared methods in terms of both evaluation indices in all patterns of $(\eta, \xi)$ in Error 2 for $p = 100$ (Table 2) and Error 4 for both $p = 12$ (Table 3) and 100 (Table 4). Focusing on the median APE, the proposed method was better in all patterns of $(\eta, \xi)$ in Error 3 for $p = 12$ (Table 4); Error 1 for $p = 12$ (Table 2), except $\eta = 0.25$; and Error 2 of $p = 12$ (Table 3), except $(\eta, \xi) = (0.75, 0.05)$. In terms of the MSE of $\beta_s$ (the second evaluation index), the proposed method had the smallest values in all patterns except $(\eta, \xi) = (0.75, 0.05)$ in Error 2 for $p = 12$ (Table 3); and Error 3 for $p = 12$ except $\eta = 0.25$. For the compared methods, the MCEN had smaller values in almost all patterns of Error 3 for $p = 100$ (Table 8) in both evaluation indices. For median APE, the MCEN was better in several patterns of Error 1 for $p = 100$ (Table 6) as well as in Error 1 for $p = 12$ (Table 2), and 100 (Table 9) for the MSE of $\beta_s$. SEN was better for several patterns of $\eta = 0.25$ of Error 1 for $p = 12$ (Table 2), for both evaluations indices as well as in Error 1 for $p = 100$ (Table 6) for the median APE. WLASSO was only better than the other methods in $\eta = 0.75$ and $\xi = 0.05$ in Error 2 for $p = 12$ (Table 3).

Next, observing in the covariate variables of Factor 2 on the median of APE, the proposed method was superior to SEN in $\eta = 0.50, 0.75$, and $1.00$ in $p = 12$ for Error 1. However, the values of MCEN and SEN were smaller than the proposed method in the same pattern for $p = 100$. In Error 3 for $p = 12$, the proposed method outperformed MCEN for most patterns of both evaluation measures, while MCEN was better for $p = 100$. In Error 1 and Error 3, the values of the proposed method and MCEN were close each other, regardless of the value of the covariate variables. The results on Error 2 and Error 4 did not vary depending on the covariate variables. As for Factor 3, parameter $\eta$, the values of all methods increased in proportion to the augment of $\eta$. Meanwhile, the values did not tend to change depending on $\xi$, Factor 4.

Finally, we see the results of the error distribution, Factor 5. In Error 1, the proposed method was almost stable for $p = 12$ in the median APE, while SEN and MCEN were better than the proposed method for $p = 100$. MCEN had the smallest MSE of $\beta_s$ for both $p = 12$ and $p = 100$. However, the value differences among the methods were slight in both the median APE and the MSE of $\beta_s$. In the case of Error 2, the situation where the error contains outliers, for both $p = 12$ and $p = 100$, the proposed method was better than the other compared methods in both the median APE and the MSE of $\beta_s$. Focusing on the differences, the value differences between the proposed method and WLASSO were closer than those between the proposed method and MCEN and SEN. In Error 3, whose error distribution follows $t$ distribution, the median APE of the proposed method was better in $(\eta, \xi) = (0.25, 0.05), (0.25, 0.10)$, and $(0.50, 0.10)$, and MCEN had better results in the other pattern of $\eta$ and $\xi$. In the MSE of $\beta_s$, the proposed method was the smallest for all for $p = 12$, except $\eta = 0.25$. For $p = 100$, MCEN was smaller than the proposed method; however, the differences were minor. Finally, in Error 4, the other heavy-tailed error distribution, the proposed method was better than all the other compared methods. For $p = 12$ in Error 4, WLASSO was also stable in terms of both the median of APE and MSE of $\beta_s$ compared to MCEN and SEN.

| $\eta$ | the median of APE (sd of APE) | | | MSE of $\beta_s$ (sd of MSE) | | |
|-------|-----------------------------|---|---|-----------------------------|---|---|
|       | $\xi$ | 0.02 | 0.05 | 0.10 | 0.02 | 0.05 | 0.10 |
| 0.25  | WMCEN | 0.706 (0.010) | 0.706 (0.010) | 0.706 (0.011) | 0.023 (0.004) | 0.023 (0.004) | 0.023 (0.004) |
|       | MCEN  | 0.712 (0.005) | 0.711 (0.014) | 0.715 (0.014) | 0.019 (0.004) | 0.017 (0.004) | 0.019 (0.003) |
|       | SEN   | 0.703 (0.010) | 0.705 (0.010) | 0.705 (0.010) | 0.018 (0.003) | 0.019 (0.004) | 0.018 (0.004) |
|       | WLASSO| 0.720 (0.012) | 0.720 (0.012) | 0.714 (0.011) | 0.035 (0.007) | 0.034 (0.007) | 0.029 (0.005) |
| 0.50  | WMCEN | 0.724 (0.013) | 0.724 (0.013) | 0.721 (0.012) | 0.035 (0.007) | 0.037 (0.007) | 0.035 (0.006) |
|       | MCEN  | 0.741 (0.034) | 0.741 (0.033) | 0.746 (0.033) | 0.029 (0.008) | 0.029 (0.008) | 0.032 (0.007) |
|       | SEN   | 0.725 (0.012) | 0.724 (0.012) | 0.724 (0.012) | 0.035 (0.006) | 0.036 (0.006) | 0.035 (0.007) |
|       | WLASSO| 0.730 (0.013) | 0.735 (0.014) | 0.734 (0.014) | 0.043 (0.008) | 0.047 (0.008) | 0.046 (0.008) |
| 0.75  | WMCEN | 0.724 (0.014) | 0.724 (0.013) | 0.736 (0.014) | 0.038 (0.008) | 0.038 (0.008) | 0.039 (0.008) |
|       | MCEN  | 0.764 (0.062) | 0.765 (0.062) | 0.766 (0.062) | 0.028 (0.009) | 0.028 (0.009) | 0.041 (0.009) |
|       | SEN   | 0.735 (0.013) | 0.734 (0.013) | 0.729 (0.013) | 0.046 (0.009) | 0.043 (0.009) | 0.041 (0.009) |
|       | WLASSO| 0.739 (0.015) | 0.739 (0.015) | 0.739 (0.015) | 0.051 (0.010) | 0.051 (0.010) | 0.051 (0.010) |
| 1.00  | WMCEN | 0.725 (0.014) | 0.725 (0.015) | 0.726 (0.014) | 0.039 (0.009) | 0.040 (0.009) | 0.052 (0.009) |
|       | MCEN  | 0.805 (0.100) | 0.803 (0.098) | 0.803 (0.099) | 0.034 (0.013) | 0.034 (0.013) | 0.034 (0.013) |
|       | SEN   | 0.735 (0.014) | 0.735 (0.015) | 0.737 (0.014) | 0.048 (0.010) | 0.045 (0.010) | 0.049 (0.010) |
|       | WLASSO| 0.740 (0.015) | 0.740 (0.015) | 0.740 (0.015) | 0.052 (0.010) | 0.052 (0.010) | 0.052 (0.010) |
### Table 3: Result of the simulation $p = 12$ for Error 2

| $\eta$ | WMCEN | MCEN | SEN | WLAAMO |
|--------|--------|------|-----|--------|
| 0.25   | 0.756 (0.011) | 0.757 (0.014) | 0.757 (0.014) | 0.025 (0.006) | 0.026 (0.008) | 0.026 (0.008) |
| 0.50   | 0.778 (0.017) | 0.779 (0.017) | 0.779 (0.018) | 0.043 (< 0.001) | 0.044 (0.010) | 0.044 (0.010) |
| 0.75   | 0.794 (0.024) | 0.833 (0.031) | 0.791 (0.021) | 0.057 (0.017) | 0.093 (0.028) | 0.054 (0.015) |
| 1.00   | 0.791 (0.026) | 0.796 (0.026) | 0.793 (0.024) | 0.056 (0.020) | 0.059 (0.024) | 0.056 (0.019) |

### Table 4: Result of the simulation $p = 12$ for Error 3

| $\eta$ | WMCEN | MCEN | SEN | WLAAMO |
|--------|--------|------|-----|--------|
| 0.25   | 1.097 (0.018) | 1.101 (0.019) | 1.100 (0.018) | 0.039 (0.008) | 0.042 (0.009) | 0.040 (0.009) |
| 0.50   | 1.128 (0.019) | 1.128 (0.019) | 1.126 (0.019) | 0.069 (0.011) | 0.069 (0.011) | 0.067 (0.010) |
| 0.75   | 1.150 (0.022) | 1.150 (0.022) | 1.150 (0.021) | 0.093 (0.015) | 0.093 (0.015) | 0.098 (0.016) |
| 1.00   | 1.159 (0.024) | 1.163 (0.024) | 1.159 (0.022) | 0.104 (0.020) | 0.108 (0.021) | 0.105 (0.019) |

### Table 5: Result of the simulation $p = 12$ for Error 4

| $\eta$ | WMCEN | MCEN | SEN | WLAAMO |
|--------|--------|------|-----|--------|
| 0.25   | 1.134 (0.033) | 1.102 (0.027) | 1.105 (0.026) | 0.072 (0.023) | 0.047 (0.013) | 0.048 (0.013) |
| 0.50   | 1.157 (0.031) | 1.157 (0.031) | 1.159 (0.031) | 0.084 (0.018) | 0.084 (0.018) | 0.084 (0.018) |
| 0.75   | 1.221 (0.041) | 1.205 (0.037) | 1.204 (0.037) | 0.137 (0.034) | 0.118 (0.025) | 0.118 (0.025) |
| 1.00   | 1.249 (0.048) | 1.249 (0.048) | 1.233 (0.043) | 0.160 (0.041) | 0.161 (0.042) | 0.142 (0.031) |
| η   | WMCEN | MCEN | SEN | WLAFFS | MSE (median of APE) | MSE (sd of APE) | β₀ (median of APE) | MSE (sd of APE) |
|-----|-------|------|-----|---------|------------------|----------------|------------------|-----------------|
| 0.20 | 0.884 (0.016) | 0.844 (0.017) | 0.844 (0.017) | 0.007 (< 0.001) | 0.007 (< 0.001) | 0.007 (< 0.001) | 0.007 (< 0.001) |
| 0.50 | 0.906 (0.033) | 0.999 (0.002) | 1.002 (0.033) | 0.014 (0.002) | 0.015 (0.002) | 0.015 (0.002) | 0.015 (0.002) |
| 0.75 | 1.078 (0.051) | 1.078 (0.051) | 1.081 (0.050) | 0.021 (0.003) | 0.021 (0.003) | 0.021 (0.003) | 0.021 (0.003) |
| 1.00 | 1.165 (0.132) | 1.166 (0.132) | 1.165 (0.131) | 0.021 (0.004) | 0.021 (0.004) | 0.021 (0.004) | 0.021 (0.004) |

Table 6: Result of the simulation p = 100 for Error 1

| η   | WMCEN | MCEN | SEN | WLAFFS | MSE (median of APE) | MSE (sd of APE) | β₀ (median of APE) | MSE (sd of APE) |
|-----|-------|------|-----|---------|------------------|----------------|------------------|-----------------|
| 0.20 | 0.871 (0.024) | 0.891 (0.024) | 0.874 (0.022) | 0.005 (< 0.001) | 0.006 (0.001) | 0.005 (0.001) | 0.005 (0.001) |
| 0.50 | 1.087 (0.049) | 1.090 (0.052) | 1.088 (0.048) | 0.016 (0.002) | 0.016 (0.002) | 0.016 (0.002) | 0.016 (0.002) |
| 0.75 | 1.285 (0.112) | 1.245 (0.095) | 1.224 (0.086) | 0.028 (0.006) | 0.026 (0.005) | 0.026 (0.005) | 0.026 (0.005) |
| 1.00 | 1.596 (0.323) | 1.600 (0.300) | 1.600 (0.305) | 0.065 (0.029) | 0.057 (0.028) | 0.057 (0.028) | 0.057 (0.028) |

Table 7: Result of the simulation p = 100 for Error 2

| η   | WMCEN | MCEN | SEN | WLAFFS | MSE (median of APE) | MSE (sd of APE) | β₀ (median of APE) | MSE (sd of APE) |
|-----|-------|------|-----|---------|------------------|----------------|------------------|-----------------|
| 0.20 | 1.293 (0.012) | 1.206 (0.024) | 1.224 (0.022) | 0.011 (0.001) | 0.006 (0.001) | 0.009 (0.001) | 0.009 (0.001) |
| 0.50 | 1.383 (0.026) | 1.386 (0.061) | 1.422 (0.060) | 0.016 (0.003) | 0.017 (0.003) | 0.018 (0.004) | 0.018 (0.004) |
| 0.75 | 1.584 (0.033) | 1.438 (0.039) | 1.405 (0.035) | 0.020 (0.002) | 0.022 (0.002) | 0.019 (0.002) | 0.019 (0.002) |
| 1.00 | 1.754 (0.085) | 1.754 (0.086) | 1.752 (0.082) | 0.049 (0.007) | 0.049 (0.007) | 0.049 (0.007) | 0.049 (0.007) |

Table 8: Result of the simulation p = 100 for Error 3
Table 9: Result of the simulation $p = 100$ for Error 4

| $\xi$ | the median of APE (sd of APE) | MSE of $\beta_s$ (sd of MSE) |
|-------|-------------------------------|-----------------------------|
|       | 0.02 0.05 0.10 | 0.02 0.05 0.10 |
| 0.25  | WMCE 1.377 (0.043) 1.378 (0.043) 1.401 (0.045) | 0.011 (0.001) 0.011 (0.001) 0.012 (0.002) |
|       | MCEN 2.473 (0.542) 2.473 (0.540) 2.472 (0.535) | 1.479 (0.455) 1.512 (0.630) 1.461 (0.968) |
|       | SEN 2.393 (0.643) 2.638 (0.789) 2.425 (0.669) | 582.884 (503.615) 670.449 (5871.294) 535.364 (4371.789) |
|       | WLASSO 3.202 (0.401) 3.202 (0.400) 3.206 (0.399) | 0.218 (0.144) 0.218 (0.144) 0.219 (0.144) |
| 0.50  | WMCE 1.664 (0.057) 1.640 (0.058) 1.630 (0.062) | 0.026 (0.003) 0.025 (0.003) 0.024 (0.003) |
|       | MCEN 3.514 (0.809) 2.895 (0.545) 2.901 (0.549) | 39.493 (243.653) 1.099 (0.496) 0.833 (5.079) |
|       | SEN 3.247 (0.812) 3.355 (0.893) 3.029 (0.726) | 730.828 (6363.460) 608.572 (5044.895) 713.699 (6253.216) |
|       | WLASSO 3.507 (0.422) 3.383 (0.397) 3.382 (0.394) | 0.303 (0.173) 0.248 (0.143) 0.249 (0.143) |
| 0.75  | WMCE 1.916 (0.111) 1.971 (0.123) 1.930 (0.110) | 0.041 (0.005) 0.045 (0.005) 0.042 (0.004) |
|       | MCEN 3.565 (0.798) 3.433 (0.722) 3.556 (0.791) | 25.040 (139.366) 18.090 (97.046) 24.118 (133.008) |
|       | SEN 3.776 (0.838) 3.756 (0.894) 3.769 (0.889) | 666.696 (5723.088) 631.024 (5330.103) 723.491 (6093.913) |
|       | WLASSO 3.978 (0.484) 3.979 (0.483) 3.980 (0.484) | 61.319 (510.994) 61.317 (510.980) 61.315 (510.968) |
| 1.00  | WMCE 2.217 (0.104) 2.261 (0.207) 2.253 (0.202) | 0.070 (0.008) 0.066 (0.009) 0.067 (0.009) |
|       | MCEN 3.807 (0.754) 3.789 (0.741) 3.773 (0.767) | 23.917 (123.683) 22.480 (114.956) 20.724 (105.953) |
|       | SEN 4.413 (0.924) 4.428 (1.017) 4.351 (0.947) | 612.964 (5383.766) 646.733 (5487.569) 688.426 (6132.423) |
|       | WLASSO 4.248 (0.506) 4.169 (0.482) 4.063 (0.463) | 61.386 (511.104) 68.686 (931) 0.405 (0.175) |

4 Real Example

In this section, we applied the proposed method to a genetic dataset named "chin07" from package lol in R software [Chin et al., 2007] to verify the method's usefulness. The dataset consists of copy number patterns and mRNA expression levels in genomic regions with candidate oncogenes for breast cancer. In recent years, with the development of computational technologies, there has been a lot of research on identifying genomic regions with candidate oncogenes through genome-wide profiling, with the goal of suppressing the expression of cancer. Altered DNA copy numbers are considered to be one of the causes of genetic abnormalities leading to disease. Identifying which DNA copy number alternation patterns affect the mRNA expression level from this dataset is key. Therefore, it is necessary to improve the accuracy of estimation along with variable selection. This dataset consists of a matrix of seven types of mRNA expressions highly relevant to breast cancer with 106 samples and a matrix of DNA copy number data for 339 regions with 106 samples. We set the mRNA expression levels as response variables and the DNA copy numbers as explanatory variables. Figure 1 draws a scatter plot, histogram, and correlations of the response variables. GI_17318566-A, Hs.500472-S, and GI_4758297-S appear to have right-tailed distribution, while GI_38505204-S has a left-tailed distribution. GI_17318560-A contains outliers. We compare the proposed method with MCEN, SEN, and WLASSO just as we did with the numerical simulation.

Figure 2 draws a scatter plot, histogram, and correlations of the response variables. GI_17318566-A, Hs.500472-S, and GI_4758297-S appear to have right-tailed distribution, while GI_38505204-S has a left-tailed distribution. GI_17318560-A contains outliers. We compare the proposed method with MCEN, SEN, and WLASSO just as we did with the numerical simulation.

Next, we describe the evaluation procedures. The subjects are randomly split into 80 samples for the training data and 26 samples for the test data. Each method is then used to estimate $\beta_s (s = 1, 2, \ldots, 7)$ with the training data, and an evaluation index is calculated by applying the estimated $\beta_s (s = 1, 2, \ldots, 7)$ to the test data. For the evaluation index, the median APE is $|\hat{y}_{s,i} - \tilde{y}_{s,i}| (i = 1, 2, \ldots, 26, s = 1, 2, \ldots, 7)$, where $\hat{y}_{s,i}$ is the predicted value, which is set in the same manner as [Wang and Li, 2009]. In this evaluation, both the response variables and explanatory variables were standardized in the training and test data in the same manner as [Mukherjee and Zhu, 2011]. A five-fold cross-validation is applied to the training data to decide the tuning parameter in all methods. Candidate $k$ is set as 2 and 3 for the proposed method and MCEN. MCEN, SEN, and WLASSO use the same R software packages as in the numerical simulation. We compare the four methods by each response variable.

The results for the response variables are plotted in Figure 2. The vertical axis describes the median APE. We compare the result of the proposed method to each compared method. First, compared to MCEN, the proposed method had good results in the response variables GI_17318566-A, GI_31543215-S, Hs.500472-S, GI_38505204-S, and GI_16950654-S. The proposed method also performed better than WLASSO for all response variables. In terms of the comparison with SEN, the proposed method had smaller evaluation values in the response variables GI_17318566-A, Hs.500472-S, GI_38505204-S, and GI_17318560-S.

The results differed for each response variable, so we discuss the results from the distribution of the response variables in Figure 1. The proposed method was superior to the other methods in the response variables that have a heavy tail such as GI_17318566-A, Hs.500472-S, GI_38505204-S, and GI_4758297-S. We focus on the shapes of the distribution
Figure 1: Scatter plot, histogram, and correlations of the response variables
of the response variables, for which the proposed method performed well. In the case of the comparison to MCEN, the proposed method was better than MCEN, regardless of the shape of the distribution of the response variables. Next, in the results compared to SEN, the proposed method had better outcomes for the variables GI_17318566-A, Hs.500472-S and GI_38505204-S, which follow heavy-tailed distributions. In the variables GI_31543215-S, GI_4758297-S, and GI_17318560-A, the proposed method was minimally inferior to SEN, which might be due to the selection of the tuning parameters on each response variable in SEN.

5 Discussion and Conclusion

In the numerical simulation, the proposed method was superior when the errors followed heavy-tailed distributions or the data contained outliers, like Error 2, Error 3, and Error 4. Especially in Error 2, which contained outliers, and Error 4, which had a strongly heavy-tailed setting, the proposed method performed much better than the other compared methods in both the median APE and MSE of $\beta_s$. This indicates that the proposed method is stable to heavy-tailed error distributions and outliers in the error. From these results, our proposed method obtained the similar results as Wang et al. [2020]. In addition, among the compared methods, WLASSO had closer results to the proposed method than MCEN and SEN in Error 2. This shows that a Wilcoxon-type regression function is effective when there are outliers. Moreover, our proposed method extended the Wilcoxon-type regression function in the framework of multivariate regression, which enabled it to take into account the correlation of response variables. Therefore, the results of the proposed method were better than WLASSO. In Error 1, where the error follows normal distribution, the proposed method was superior to the other methods in $p = 12$. This result is consistent with the trend of results in the error with normal distribution in Wang et al. [2020]. However, SEN and MCEN are had better results for $p = 100$ due to the assumption of a normal distribution for the residuals. When focusing on value differences, the values of the proposed method were close to those of the compared methods in both $p = 12$ and $p = 100$. Therefore, the proposed method can retain comparable estimation accuracy to the compared methods in situations where the error distribution follows a normal distribution, regardless of the number of variables.

In the real data application, the proposed method was better than the other methods in the responses GI_17318566-A, Hs.500472-S, GI_38505204-S, and GI_4758297-S, which followed heavy-tailed distribution. The proposed method showed the stability of heavy-tailed situations in real data as well. Meanwhile, in the variables GI_4758297-S and GI_17318560-A, the proposed method was minimally inferior to SEN. The difference of the median APE between the proposed method and SEN was not very broad. This is considered to be due to the fact that SEN selects the tuning parameters on each response variable in SEN.
parameters for each response variable, while the proposed method selected single tuning parameters for all response variables in this study.

We showed the efficacy of the proposed method through the numerical simulation and real data application. From these results, this method will be expected to provide high estimation accuracy when applied to actual multivariate data which includes heavy-tailed distributions and contains outliers. To improve our proposed method, four points need to be considered. First, in this study, we set common tuning parameters for all response variables. We expect that the estimation accuracy of the method would enhance more if different tuning parameters are set depending on each response variable. Secondly, we showed that the Wilcoxon-type regression is more efficient with respect to the tailed distribution or outliers of the response. However, Wilcoxon-type regression is not very robust against the outliers in explanatory variables. To overcome this, the weight of the Wilcoxon-type regression can be set, which has been proposals in several studies [Naranjo and Hettmansperger, 1994; Sievers, 1983; Wang and Li, 2009]. This will need further consideration in situations where the explanatory variables contain outliers. Third, our extension of the Wilcoxon-type regression to multivariate regression is formulated differently than the method proposed by Weihua [2010]. Thus, we need to compare the two methods from various perspectives. Finally, we extended the Wilcoxon-type regression in the framework of multivariate regression. However, several robust reduced-rank regressions for multivariate regression have also been proposed [Chao et al., 2021; Ding et al., 2021; Wang and Karunamuni, 2022; Tan et al., 2022; Wang and Li, 2009]. This will need further consideration in situations where the explanatory variables contain outliers. The difference between our proposed method and these robust reduced-rank methods is that the proposed method has a clustering term to group the fitted values of the response variables, which allows the correlation among the response variables to be considered. Further consideration of this will be needed.

References

L. Breiman and J. H. Friedman. Predicting multivariate responses in multiple linear regression. *Journal of the Royal Statistical Society: Series B* (Statistical Methodology), 59(1):3–54, 1997.

A. J. Rothman, E. Levina, and Z. Ji. Sparse multivariate regression with covariance estimation. *Journal of Computational and Graphical Statistics*, 19(4):947–962, 2010.

J. Peng, J. Zhu, A. Bergamaschi, W. Han, D.Y. Noh, J.R. Pollack, and P. Wang. Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer. *The annals of applied statistics*, 4(1):53, 2010.

S. Kim and E.P. Xing. Tree-guided group lasso for multi-response regression with structured sparsity, with an application to eQTL mapping. *The Annals of Applied Statistics*, 6(3):1095–1117, 2012.

Y. Chen, R. Iyengar, and G. Iyengar. Modeling multimodal continuous heterogeneity in conjoint analysis—a sparse learning approach. *Marketing Science*, 36(1):140–156, 2017.

R. D. Cook, B. Li, and F. Chiaromonte. Envelope models for parsimonious and efficient multivariate linear regression. *Statistica Sinica*, pages 927–960, 2010.

R.D. Cook and X. Zhang. Foundations for envelope models and methods. *Journal of the American Statistical Association*, 110(510):599–611, 2015.

Q. Sun, H. Zhu, Y. Liu, and J.G. Ibrahim. Sprem: sparse projection regression model for high-dimensional linear regression. *Journal of the American Statistical Association*, 110(509):289–302, 2015.

B. S. Price and B. Sherwood. A cluster elastic net for multivariate regression. *Journal of Machine Learning Research*, pages 1–30, 2018.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

E. W. Forgy. Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *biometrics*, 21:768–769, 1965.

P.J. Huber. Robust statistics. In *International encyclopedia of statistical science*, pages 1248–1251. Springer, 2011.

J. Fan, Q. Li, and Y. Wang. Estimation of high dimensional mean regression in the absence of symmetry and light tail assumptions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(1):247–265, 2017.

P.L. Loh. Statistical consistency and asymptotic normality for high-dimensional robust m-estimators. *The Annals of Statistics*, 45(2):866–896, 2017.

Q. Sun, W.X. Zhou, and J. Fan. Adaptive huber regression. *Journal of the American Statistical Association*, 2019.

A.C. Lozano, N. Meinshausen, and E. Yang. Minimum Distance Lasso for robust high-dimensional regression. *Electronic Journal of Statistics*, 10(1):1296 – 1340, 2016.
L. Wang, Y. Wu, and R. Li. Robust Variable Selection With Exponential Squared Loss. *Journal of American Statistical Association*, 108(1):632 – 643, 2012a.

M. Avella-Medina and E. Ronchetti. Robust and consistent variable selection in high-dimensional generalized linear models. *Biometrika*, 105:31–44, 2018.

A. Prasad, A.S. Sai, S. Balakrishnan, and P. Ravikumar. Robust estimation via robust gradient estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82, 2020.

A. Belloni, V. Chernozhukov, and L. Wang. Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806, 2011.

J. Bradic, J. Fan, and W. Wang. Penalized composite quasi-likelihood for ultrahigh dimensional variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(3):325–349, 2011.

L. Wang, Y. Wu, and R. Li. Quantile regression for analyzing heterogeneity in ultra-high dimension. *Journal of the American Statistical Association*, 107(497):214–222, 2012b.

L. Wang. The l1 penalized lad estimator for high dimensional linear regression. *Journal of Multivariate Analysis*, 120:135–151, 2013.

J. Fan, Y. Fan, and E. Barut. Adaptive robust variable selection. *Annals of statistics*, 42 1:324–351, 2014.

T. Sun and C.H. Zhang. Scaled sparse linear regression. *Biometrika*, 99(4):879–898, 2012.

P.J. Huber. Robust estimation of a location parameter. *Annals of Mathematical Statistics*, 35:492–518, 1964.

H. Wang, G. Li, and G. Jiang. Robust regression shrinkage and consistent variable selection through the lad-lasso. *Journal of Business & Economic Statistics*, 25(3):347–355, 2007.

H. Zou and M. Yuan. Composite quantile regression and the oracle model selection theory. *The Annals of Statistics*, 36 (3):1108–1126, 2008.

J. Jureckova. Nonparametric estimate of regression coefficients. *The Annals of Mathematical Statistics*, pages 1328–1338, 1971.

L. A. Jaeckel. Estimating regression coefficients by minimizing the dispersion of the residuals. *The Annals of Mathematical Statistics*, pages 1449–1458, 1972.

T.P. Hettmansperger and J.W. McKean. A robust alternative based on ranks to least squares in analyzing linear models. *Technometrics*, 19(3):275–284, 1977.

T. P. Hettmansperger and J. W. McKean. *Robust nonparametric statistical methods*. Arnold, 1998.

L. Wang, B. Peng, J. Bradic, R. Li, and Y. Wu. A tuning-free robust and efficient approach to high-dimensional regression. *Journal of the American Statistical Association*, 115(532):1700–1714, 2020.

L. Wang and R. Li. Weighted wilcoxon-type smoothly clipped absolute deviation method. *Biometrics*, 65(2):564–571, 2009.

D. R. Hunter and K. Lange. A tutorial on mm algorithms. *The American Statistician*, 58(1):30–37, 2004.

T. P. Hettmansperger and J. W. McKeans. Statistical inference based on ranks. *Psychometrika*, 43:69–79, 1978.

D. R. Hunter and R. Li. Variable selection using mm algorithm. *Annals of statistics*, 33(4):1617, 2005.

D. Yu, J. H. Won, T. Lee, J. Lim, and S. Yoon. High-dimensional fused lasso regression using majorization-minimization and parallel processing. *Journal of Computational and Graphical Statistics*, 24(1):121–153, 2015.

F.W. Young. Quantitative analysis of qualitative data. *Psychometrika*, 46:357–388, 1981.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B (statistical methodology)*, 67(2):301–320, 2005.

B. Sherwood and B. Price. *mcen: Multivariate Cluster Elastic Net*, 2020. URL https://CRAN.R-project.org/package=mcen R package version 1.2.

J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1, 2008.

S. Li and Y. Qin. *MTE: Maximum Tangent Likelihood Estimation for Linear Regression*, 2021. URL https://CRAN.R-project.org/package=MTE R package version 1.0.1.

S. F. Chin, A. E. Teschendorff, J. C. Marioni, Y. Wang, N. L. Barbosa-Morais, N. P. Thorne, J. L. Costa, S. E. Pinder, M. A. Van de Wiel, A. R Green, et al. High-resolution acgh and expression profiling identifies a novel genomic subtype of er negative breast cancer. *Genome biology*, 8(10):1–17, 2007.
A. Mukherjee and J. Zhu. Reduced rank ridge regression and its kernel extensions. *Statistical analysis and data mining: the ASA data science journal*, 4(6):612–622, 2011.

J.D. Naranjo and T.P. Hettmansperger. Bounded influence rank regression. *Journal of the Royal Statistical Society: Series B (Methodological)*, 56(1):209–220, 1994.

G.L. Sievers. A weighted dispersion function for estimation in linear models. *Communications in Statistics-Theory and Methods*, 12(10):1161–1179, 1983.

Z. Weihua. A multivariate wilcoxon regression estimate. *Journal of Nonparametric Statistics*, 22(7):859–877, 2010.

S. K. Chao, W. K. Härdle, and M. Yuan. Factorisable multitask quantile regression. *Econometric Theory*, 37(4):794–816, 2021.

H. Ding, S. Qin, Y. Wu, and Y. Wu. Asymptotic properties on high-dimensional multivariate regression m-estimation. *Journal of Multivariate Analysis*, 183:104730, 2021.

Y. Wang and R. J. Karunamuni. High-dimensional robust regression with lq-loss functions. *Computational Statistics and Data Analysis*, page 107567, 2022.

K. M. Tan, Q. Sun, and D. Witten. Sparse reduced rank huber regression in high dimensions. *Journal of the American Statistical Association*, pages 1–11, 2022.

A. Mishra and C. L. Müller. Robust regression with compositional covariates. *Computational Statistics and Data Analysis*, 165:107315, 2022.

W. Zhao, H. Lian, and S. Ma. Robust reduced-rank modeling via rank regression. *Journal of Statistical Planning and Inference*, 180:1–12, 2017.

Y. She and K. Chen. Robust reduced-rank regression. *Biometrika*, 104(3):633–647, 2017.