

Confidence intervals for high-dimensional multi-task regression

Yuanli Ma¹, Yang Li² , and Jianjun Xu²

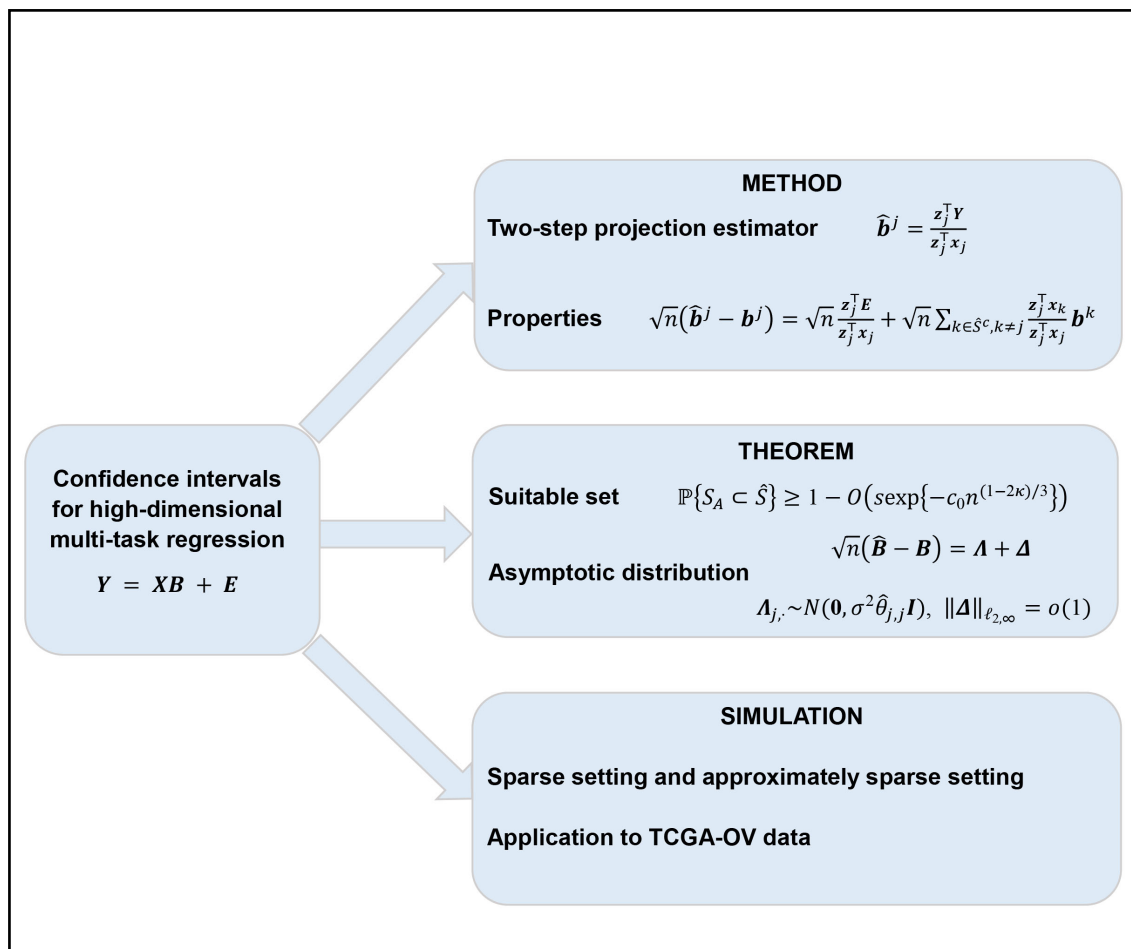
¹*School of Data Science, University of Science and Technology of China, Hefei 230026, China;*

²*International Institute of Finance, School of Management, University of Science and Technology of China, Hefei 230026, China*

Correspondence: Yang Li, E-mail: tjly@mail.ustc.edu.cn

© 2023 The Author(s). This is an open access article under the CC BY-NC-ND 4.0 license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

Graphical abstract




Statistical inference for coefficient matrix in high-dimensional multi-task regression.

Public summary


- We propose a two-step projection estimator for statistical inference in high-dimensional multi-task learning problems.
- We establish the asymptotic properties of the proposed estimator.
- The performance of our method is presented through simulation studies and a TCGA-OV dataset.

Confidence intervals for high-dimensional multi-task regression

Yuanli Ma¹, Yang Li² , and Jianjun Xu²

¹School of Data Science, University of Science and Technology of China, Hefei 230026, China;

²International Institute of Finance, School of Management, University of Science and Technology of China, Hefei 230026, China

 Correspondence: Yang Li, E-mail: tjly@mail.ustc.edu.cn

© 2023 The Author(s). This is an open access article under the CC BY-NC-ND 4.0 license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).



Cite This: *JUSTC*, 2023, 53(4): 0403 (9pp)



Read Online

Abstract: Regression problems among multiple responses and predictors have been widely employed in many applications, such as biomedical sciences and economics. In this paper, we focus on statistical inference for the unknown coefficient matrix in high-dimensional multi-task learning problems. The new statistic is constructed in a row-wise manner based on a two-step projection technique, which improves the inference efficiency by removing the impacts of important signals. Based on the established asymptotic normality for the proposed two-step projection estimator (TPE), we generate corresponding confidence intervals for all components of the unknown coefficient matrix. The performance of the proposed method is presented through simulation studies and a real data analysis.

Keywords: statistical inference; confidence interval; two-step projection; bias correction; feature screening

CLC number: O212; TP391

Document code: A

1 Introduction

With the advent of massive data in recent years, great attention has been given to identifying relationships among multiple responses and predictors in various applications, including multi-task learning in machine learning^[1–3], imaging genetics^[4,5] and genetic association^[6,7]. Specifically, in cancer genomic studies, some miRNAs are known to regulate protein expression of various genes in cellular processes and their dysregulation plays a crucial role in human cancer. Hence, investigating the relationship between miRNA expression and protein expression is of great importance for human cancer diagnosis^[8–9]. A standard approach for quantifying these relationships is to perform a univariate regression model for each response separately via the least squares estimation. Although it is easy to implement, this method ignores the dependence information among response variables, and it is also not applicable to high-dimensional cases. Thus, it is necessary to develop statistical tools that account for the dependence structure of responses in multi-response regression under high-dimensional settings.

An enormous effort has been mounted for variable selection and coefficient estimation in high-dimensional multi-response regression. Among them, one popular way is to consider the reduced rank regression model, including some foundational works^[10–12]. Based on this framework, several methods that apply a latent factor point of view have been proposed to estimate the unknown coefficient matrix^[13,14]. Another class of methods relies on different kinds of regularization^[15–18]. Specifically, a line of research for regularization methods focuses on some structural prior knowledge of the coefficient matrix, that is, the set of variables is assumed to be structured into several groups. Please refer to Refs. [19–21] for more details about this kind of method.

To further assign uncertainty, some important progress has been made in high-dimensional multi-response settings. For instance, Greenlaw et al.^[5] developed a hierarchical Bayesian model and constructed confidence intervals for the unknown coefficient matrix. More recently, by generalizing low-dimensional projection estimation (LDPE)^[22] in the univariate response case, Chevalier et al.^[23] proposed the desparsified multi-task lasso method and applied it to source imaging. However, when deriving the asymptotic distributions of estimators, this approach requires the number of nonzero rows of coefficient matrix to be $s = o(\sqrt{n}/\log(p))$ for p covariates and n samples. To further alleviate this requirement, we attempt to use the two-step projection technique proposed by Li et al.^[24], which treats important variables and others differently and can greatly improve the inference efficiency.

In this paper, we aim to develop a new methodology for statistical inference in the setting of high-dimensional multi-task regression. By taking group structures of the unknown coefficient matrix into consideration, the proposed estimator is constructed in a row-wise manner based on the two-step projection technique, which enjoys the benefit of reducing the estimation bias induced by these important signals. Under suitable conditions, we establish the asymptotic normality of the proposed two-step projection estimator along with corresponding confidence intervals for all components of the unknown coefficient matrix. Moreover, the satisfactory numerical performance of the proposed method strongly supports the theoretical results.

The rest of the paper is organized as follows. Section 2 presents the model setting and the new inference procedure for high-dimensional multi-response regression. Theoretical properties are established in Section 3. Numerical results and a real data analysis are provided in Sections 4 and 5, respectively. We conclude this work and possible future work in

Section 6. The proofs of main theoretical properties are delegated to Appendix.

Notations. For any matrix $\mathbf{B} = (\mathbf{b}^{1\top}, \dots, \mathbf{b}^{p\top})^\top = (\mathbf{b}_1, \dots, \mathbf{b}_q) = (b_{i,j}) \in \mathbb{R}^{p \times q}$, denote by \mathbf{b}^i and \mathbf{b}_j its i th row and j th column, respectively. We use $\|\mathbf{B}\| = \left(\sum_{i=1}^p \sum_{j=1}^q b_{i,j}^2 \right)^{1/2}$ to denote the

Frobenius norm for matrix \mathbf{B} . Given any K , \mathbf{B}_K denotes the submatrix of \mathbf{B} consisting of columns in K . We write

$\|\mathbf{B}\|_{\ell_{2,1}} = \sum_{i=1}^p \|\mathbf{b}^i\|$ and $\|\mathbf{B}\|_{\ell_{2,\infty}} = \max_{1 \leq i \leq p} \|\mathbf{b}^i\|$, where $\|\mathbf{b}^i\| =$

$\left(\sum_{j=1}^q |b_{i,j}|^2 \right)^{1/2}$ denotes the Euclidean norm for the vector \mathbf{b}^i . De-

note by $\text{supp}(\mathbf{B}) = \{i \in \{1, \dots, p\} : \mathbf{b}^i \neq \mathbf{0}\}$ the nonzero rows of \mathbf{B} with size $|\text{supp}(\mathbf{B})|$.

2 Inference procedure via two-step projection estimator

2.1 Model setting

Consider the multi-task learning problem with a high-dimensional multi-response linear regression model

$$\mathbf{y} = \mathbf{B}^\top \mathbf{x} + \mathbf{e}, \tag{1}$$

where $\mathbf{y} = (y_1, \dots, y_q)^\top$ is a q -dimensional response vector, $\mathbf{B} = (\mathbf{b}^{1\top}, \dots, \mathbf{b}^{p\top})^\top = (\mathbf{b}_1, \dots, \mathbf{b}_q) = (b_{i,j}) \in \mathbb{R}^{p \times q}$ is the unknown coefficient matrix, $\mathbf{x} = (x_1, \dots, x_p)^\top$ is the p -dimensional covariate vector, and \mathbf{e} is the q -dimensional error vector, which is independent of \mathbf{x} . Following Li et al. [25], the dimension p is allowed to be much larger than the sample size n , and the dimension q is regarded as a fixed number in this paper.

Suppose we have n independent observations $(\mathbf{x}^i, \mathbf{y}^i)_{i=1}^n$ from (\mathbf{x}, \mathbf{y}) in model (1). Using the matrix notation, the multi-response linear regression model (1) can be rewritten as

$$\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{E}, \tag{2}$$

where $\mathbf{Y} = (\mathbf{y}_1, \dots, \mathbf{y}_q) \in \mathbb{R}^{n \times q}$ is the response matrix, $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_p) \in \mathbb{R}^{n \times p}$ is the design matrix, and $\mathbf{E} = (\mathbf{e}_1, \dots, \mathbf{e}_q) = (e_{i,j}) \in \mathbb{R}^{n \times q}$ is the random error matrix that is independent of \mathbf{X} . Without loss of generality, we assume that $e_{i,j}$ s are independent and identically distributed random variables with mean zero and variance σ^2 .

We aim to construct confidence intervals for the coefficients in \mathbf{B} . Different from the classical assumption that \mathbf{B} is row-sparse, we allow \mathbf{B} to have a more complex sparsity structure. More specifically, we first establish the relationship between the distance correlation [25, 26] and sparsity structure. Define the population quantity

$$\omega_i = \text{dcorr}^2(x_i, \mathbf{y})$$

with $1 \leq i \leq p$ for the effects caused by \mathbf{b}^i . Here, the distance correlation $\text{dcorr}(\mathbf{u}, \mathbf{v})$ between two random vectors $\mathbf{u} \in \mathbb{R}^{d_u}$ and $\mathbf{v} \in \mathbb{R}^{d_v}$ is defined as

$$\text{dcorr}(\mathbf{u}, \mathbf{v}) = \frac{\text{dcov}(\mathbf{u}, \mathbf{v})}{\sqrt{\text{dcov}(\mathbf{u}, \mathbf{u})\text{dcov}(\mathbf{v}, \mathbf{v})}}$$

with the distance covariance defined as

$$\text{dcov}^2(\mathbf{u}, \mathbf{v}) = \frac{1}{c_{d_u} c_{d_v}} \int_{\mathbb{R}^{d_u+d_v}} \frac{|f_{\mathbf{u},\mathbf{v}}(\mathbf{t}, \mathbf{s}) - f_{\mathbf{u}}(\mathbf{t})f_{\mathbf{v}}(\mathbf{s})|^2}{\|\mathbf{t}\|^{1+d_u} \|\mathbf{s}\|^{1+d_v}} d\mathbf{t}d\mathbf{s},$$

where the constant $c_d = \frac{\pi^{(1+d)/2}}{\Gamma((1+d)/2)}$ for $d = d_u, d_v$, and $f_{\mathbf{u},\mathbf{v}}(\mathbf{t}, \mathbf{s})$, $f_{\mathbf{u}}(\mathbf{t})$ and $f_{\mathbf{v}}(\mathbf{s})$ are the characteristic functions of (\mathbf{u}, \mathbf{v}) , \mathbf{u} and \mathbf{v} , respectively. Please refer to Ref. [26] for more details about the distance correlation.

Because $\text{dcorr}(\mathbf{u}, \mathbf{v}) = 0$ if and only if \mathbf{u} and \mathbf{v} are independent, we establish the following relationship between the population quantity ω_i and the structure of \mathbf{B} :

$$\omega_i = 0 \Leftrightarrow \mathbf{b}^i = \mathbf{0}, \quad 1 \leq i \leq p.$$

Similar to Li et al. [25], we regard x_i as an important predictor if $\omega_i \geq cn^{-\kappa}$, where $c > 0$ and $0 \leq \kappa < 1/2$ are some constants. Moreover, we define S_A as follows to contain the indices of all important predictors:

$$S_A = \{i \in \{1, \dots, p\} : \omega_i \geq cn^{-\kappa}\}.$$

The indices of the rest of the unimportant predictors are collected by S_A^c . Correspondingly, any \mathbf{b}^i with $i \in S_A$ is regarded as an important signal in the form of a vector, while any \mathbf{b}^i with $i \in S_A^c$ is treated as a weak signal.

2.2 Two-step projection estimator

By borrowing ideas from Li et al. [24], the proposed estimator is constructed in a row-wise manner, i.e.,

$$\hat{\mathbf{b}}^j = \frac{\mathbf{z}_j^\top \mathbf{Y}}{\mathbf{z}_j^\top \mathbf{x}_j} \tag{3}$$

for any $j \in \{1, \dots, p\}$, where the two-step projection residual vector \mathbf{z}_j is defined by the following two-step procedure:

Step 1: Given a prescreened set \hat{S} for those identifiable signals in \mathbf{B} , we obtain the following residual vector by an exact orthogonalization of \mathbf{x}_k against $\mathbf{X}_{\hat{S} \setminus \{j\}}$:

$$\boldsymbol{\psi}_k^{(j)} = (\mathbf{I}_{n \times n} - \mathbf{P}_{\hat{S} \setminus \{j\}}) \mathbf{x}_k, \tag{4}$$

where $\mathbf{P}_{\hat{S} \setminus \{j\}} = \mathbf{X}_{\hat{S} \setminus \{j\}} (\mathbf{X}_{\hat{S} \setminus \{j\}}^\top \mathbf{X}_{\hat{S} \setminus \{j\}})^{-1} \mathbf{X}_{\hat{S} \setminus \{j\}}^\top$ is the projection matrix of the column space of $\mathbf{X}_{\hat{S} \setminus \{j\}}$.

Step 2: Then, \mathbf{z}_j is constructed by the residual of lasso regression of $\boldsymbol{\psi}_j^{(j)}$ against $\boldsymbol{\psi}_{\hat{S}^c \setminus \{j\}}^{(j)}$. That is,

$$\mathbf{z}_j = \boldsymbol{\psi}_j^{(j)} - \boldsymbol{\psi}_{\hat{S}^c \setminus \{j\}}^{(j)} \hat{\mathbf{v}}_j(\lambda_j), \tag{5}$$

where $\boldsymbol{\psi}_{\hat{S}^c \setminus \{j\}}^{(j)}$ is the matrix composed of column vectors $\boldsymbol{\psi}_k^{(j)}$ for $k \in \hat{S}^c \setminus \{j\}$, and $\hat{\mathbf{v}}_j(\lambda_j)$ is the lasso estimator depending on the regularization parameter λ_j .

Based on the above two-step strategy, the two-step projection residual vector \mathbf{z}_j satisfies the following two properties:

(a) It is strictly orthogonal to $\mathbf{X}_{\hat{S} \setminus \{j\}}$ consisting of important covariates.

(b) It is relaxed orthogonally to $\mathbf{X}_{\hat{S}^c \setminus \{j\}}$ consisting of other covariates.

This kind of hybrid orthogonalization brings the benefit of reducing the estimation bias of $\hat{\mathbf{b}}^j$. To see this, plugging model (2) to (3) yields

$$\sqrt{n}(\hat{\mathbf{b}}^j - \mathbf{b}^j) = \sqrt{n} \frac{\mathbf{z}_j^\top \mathbf{E}}{\mathbf{z}_j^\top \mathbf{x}_j} + \sqrt{n} \sum_{k \neq j} \frac{\mathbf{z}_j^\top \mathbf{x}_k}{\mathbf{z}_j^\top \mathbf{x}_j} \mathbf{b}^k.$$

Since property (a) shows that

$$\mathbf{z}_j^\top \mathbf{x}_k = 0 \text{ for } k \in \hat{S} \setminus \{j\},$$

the above equality can be further rewritten as

$$\sqrt{n}(\hat{\mathbf{b}}^j - \mathbf{b}^j) = \sqrt{n} \frac{\mathbf{z}_j^\top \mathbf{E}}{\mathbf{z}_j^\top \mathbf{x}_j} + \sqrt{n} \sum_{k \in \hat{S}^c, k \neq j} \frac{\mathbf{z}_j^\top \mathbf{x}_k}{\mathbf{z}_j^\top \mathbf{x}_j} \mathbf{b}^k.$$

It is easy to see from the above that the influence generated by these identifiable signals in $\hat{S} \setminus \{j\}$ is eliminated from the bias term of the estimation error.

Denote by $\tau_j = \|\mathbf{z}_j\|/|\mathbf{z}_j^\top \mathbf{x}_j|$. Then the confidence interval of $b_{j,k}$ is constructed by

$$[\hat{b}_{j,k} - \Phi^{-1}(1 - \alpha/2)\hat{\sigma}\tau_j, \hat{b}_{j,k} + \Phi^{-1}(1 - \alpha/2)\hat{\sigma}\tau_j] \quad (6)$$

for any $j = 1, \dots, p$ and $k = 1, \dots, q$, where Φ denotes the standard normal distribution function, α is the significance level, and $\hat{\sigma}$ is a consistent estimator of σ . Following Chevalier et al.^[23] and Reid et al.^[27], in this paper, we suggest the cross-validation-based variance estimator

$$\hat{\sigma}^2 = \text{median}(\{\hat{\sigma}_t^2\}_{t \in \{1, \dots, q\}}).$$

To be specific, for $t = 1, \dots, q$,

$$\hat{\sigma}_t^2 = \|\hat{\mathbf{e}}_t\|^2/(n - \hat{s}),$$

where $\hat{\mathbf{e}}_t$ is the t th column of $\hat{\mathbf{E}} = \mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}_{\text{CV}}$ with $\hat{\mathbf{B}}_{\text{CV}}$ being the multivariate group lasso estimator tuned by cross-validation, and $\hat{s} = |\text{supp}(\hat{\mathbf{B}}_{\text{CV}})|$ means the number of nonzero rows of $\hat{\mathbf{B}}_{\text{CV}}$.

Note that a prescreened set \hat{S} for those identifiable signals in \mathbf{B} is necessary for the proposed method. In this paper, we suggest utilizing DC-SIS^[25] to obtain a suitable prescreened set \hat{S} and we will provide the sure screening property of DC-SIS in Proposition 1. In conclusion, the proposed method TPE is summarized in Algorithm 1. However, we cannot guarantee that all the truly important signals are retained in practice. In view of this potential scenario, we alternatively propose a variant two-step estimator based on the self-bias correction idea.

Given a preliminary estimate such as the multivariate group lasso estimate $\hat{\mathbf{B}}_0 = (\hat{\mathbf{b}}_0^{1\top}, \dots, \hat{\mathbf{b}}_0^{p\top})^\top$, the variant two-step estimator $\hat{\mathbf{B}}_v = (\hat{\mathbf{b}}_v^{1\top}, \dots, \hat{\mathbf{b}}_v^{p\top})^\top$ can be defined through each row as

$$\hat{\mathbf{b}}_v^j = \hat{\mathbf{b}}_0^j + \frac{\mathbf{z}_j^\top (\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}_0)}{\mathbf{z}_j^\top \mathbf{x}_j}, \quad 1 \leq j \leq p.$$

Some simple algebra shows that

$$\sqrt{n}(\hat{\mathbf{b}}_v^j - \mathbf{b}^j) = \sqrt{n} \frac{\mathbf{z}_j^\top \mathbf{E}}{\mathbf{z}_j^\top \mathbf{x}_j} + \sqrt{n} \sum_{k \in \hat{S}^c, k \neq j} \frac{\mathbf{z}_j^\top \mathbf{x}_k}{\mathbf{z}_j^\top \mathbf{x}_j} (\mathbf{b}^k - \hat{\mathbf{b}}_0^k).$$

By introducing the preliminary estimate, we can see that

$$\sqrt{n} \left\| \sum_{k \in \hat{S}^c, k \neq j} \frac{\mathbf{z}_j^\top \mathbf{x}_k}{\mathbf{z}_j^\top \mathbf{x}_j} (\mathbf{b}^k - \hat{\mathbf{b}}_0^k) \right\| \leq \sqrt{n} \left(\max_{k \neq j} \left| \frac{\mathbf{z}_j^\top \mathbf{x}_k}{\mathbf{z}_j^\top \mathbf{x}_j} \right| \right) \sum_{k \in \hat{S}^c, k \neq j} \|\mathbf{b}^k - \hat{\mathbf{b}}_0^k\|.$$

If some truly important features are omitted in the prescreening step, the variant procedure can be a reliable choice since the magnitude of the bias term depends on $\sum_{k \in \hat{S}^c, k \neq j} \|\mathbf{b}^k - \hat{\mathbf{b}}_0^k\|$

instead of the diverging term $\sum_{k \in \hat{S}^c, k \neq j} \|\mathbf{b}^k\|$.

Algorithm 1 TPE algorithm

Require: $\mathbf{X} \in \mathbb{R}^{n \times p}$, $\mathbf{Y} \in \mathbb{R}^{n \times q}$, a prescreened set \hat{S} , a significance level α ;

$$\hat{\mathbf{B}}_{\text{CV}} \leftarrow \arg \min \left\{ \frac{\|\mathbf{Y} - \mathbf{X}\mathbf{B}\|^2}{2n} + \lambda \|\mathbf{B}\|_{\ell_{2,1}} \right\};$$

$$\hat{\mathbf{E}} \leftarrow \mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}_{\text{CV}}$$

For $t \in \{1, \dots, q\}$, **do**

$$\hat{s} \leftarrow |\text{supp}(\hat{\mathbf{B}}_{\text{CV}})|;$$

$$\hat{\sigma}_t^2 \leftarrow \|\hat{\mathbf{e}}_t\|^2/(n - \hat{s})$$

End for

$$\hat{\sigma}^2 \leftarrow \text{median}(\{\hat{\sigma}_t^2\}_{t \in \{1, \dots, q\}})$$

For $j \in \{1, \dots, p\}$, **do**

$\mathbf{z}_j \leftarrow$ a two-step procedure described in Eqs. (4) and (5);

$$\hat{\mathbf{b}}^j \leftarrow \frac{\mathbf{z}_j^\top \mathbf{Y}}{\mathbf{z}_j^\top \mathbf{x}_j};$$

$$\tau_j \leftarrow \frac{\|\mathbf{z}_j\|}{|\mathbf{z}_j^\top \mathbf{x}_j|};$$

$$CI_{j,k} \leftarrow \left[\hat{b}_{j,k} - \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\hat{\sigma}\tau_j, \hat{b}_{j,k} + \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\hat{\sigma}\tau_j \right]$$

End for

Ensure: $CI_{j,k}$ for $j = 1, \dots, p$ and $k = 1, \dots, q$

3 Theoretical properties

In this section, we provide statistical properties for the proposed method TPE. First, we need to clarify some conditions on the model.

Condition 1. The rows of \mathbf{X} are independent and identically distributed (i.i.d.) from $N(\mathbf{0}, \boldsymbol{\Sigma}_X)$, and the eigenvalues of $\boldsymbol{\Theta} = \boldsymbol{\Sigma}_X^{-1} = (\theta_{i,j})$ are bounded within the interval $[1/L, L]$ for some $L \geq 1$.

Condition 2. $s^* = \max_{1 \leq j \leq p} s_j = o(n/\log(p))$, where $s_j = |\{k \in \{1, \dots, p\} : k \neq j, \theta_{j,k} \neq 0\}|$ is the sparsity with respect to rows of $\boldsymbol{\Theta}$.

Condition 3. $s = o(\max\{n/\log(p), n/s^*\})$ with $s = |S_A|$, and $\sum_{k \in S_A^c} \|\mathbf{b}^k\| = o(1/\sqrt{\log(p)})$.

Conditions 1 and 2 are the same as those in Ref. [24], which provide theoretical guarantees for estimation and prediction consistency in the two-step procedure. The first part of Condition 1 is a common Gaussian assumption, which can be relaxed to a general case such as sub-Gaussian. The second part of Condition 1 assumes that the eigenvalues of $\boldsymbol{\Theta}$ are well bounded from below and above, which is used for characterizing the identifiability of the design matrix $\boldsymbol{\Psi}_{S^c \setminus \{j\}}^{(j)}$ in Eq. (5) based on the bounded sign-restricted cone invertibility factor^[28]. Condition 2 imposes a typical constraint on the maximum column sparsity of $\boldsymbol{\Theta}$, which is needed to guarantee consistent estimation in the two-step procedure.

Condition 3 entails the main contributions of the proposed method. The first part of this condition allows the number of identifiable signals $s = o(\max\{n/\log(p), n/s^*\})$, which is much weaker than $o(\sqrt{n}/\log(p))$ in Ref. [23]. Moreover, the order of s can be much larger than $o(n/\log(p))$ if $s^* \ll \log(p)$. The second part of Condition 3 imposes a constraint on the weak signals in S_A^c , which is used to guarantee that the influence on

the weak signals cannot break the inference procedure. Next, the following definition characterizes the theoretical properties of a suitable prescreened set \hat{S} .

Definition 1 (suitable set). \hat{S} is called a suitable prescreened set if it satisfies: (a) \hat{S} is independent of (\mathbf{X}, \mathbf{Y}) ; (b) with probability at least $1 - \epsilon_{n,p}$, $S_A \subset \hat{S}$ and $|\hat{S}| = O(s)$, where $\epsilon_{n,p}$ is asymptotically vanishing.

Definition 1 is similar to the definition of an acceptable set in Ref. [24]. The first part of this definition is applied to eliminate the dependence between \hat{S} and the proposed estimator, which can be achieved by the common sample splitting technique. The second part of this definition assumes the sure screening property of \hat{S} , which can be justified by the following proposition.

Proposition 1. Under Condition 1, there exists a constant $c_0 > 0$ such that

$$\mathbb{P}\{S_A \subset \hat{S}\} \geq 1 - O(\text{sexp}\{-c_0 n^{(1-2\kappa)/3}\}),$$

where \hat{S} is obtained by the sure independence screening procedure based on the distance correlation[25].

In what follows, we present the main theoretical results of the proposed approach.

Theorem 1. Part a: Assume that \hat{S} satisfies Definition 1. The proposed two-step projection estimator satisfies

$$\sqrt{n}(\hat{\mathbf{B}} - \mathbf{B}) = \mathbf{A} + \mathbf{A},$$

$$\mathbf{A}_{j\cdot} \sim N(\mathbf{0}, \sigma^2 \hat{\theta}_{j,j} \mathbf{I}) \quad \text{with} \quad \hat{\theta}_{j,j} = n\tau_j^2$$

for any $j = 1, \dots, p$, where $\mathbf{A}_{j\cdot}$ denotes the j th row of \mathbf{A} .

Part b: Further assume that Conditions 1–3 hold. For some constants $\epsilon > 0$ and $\delta \geq 1$, let $\lambda_j = (1 + \epsilon) \sqrt{2\delta \log(p)/(n\theta_{j,j})}$ with $j = 1, \dots, p$, where λ_j is the regularization parameter in Eq. (5). Then, for any $j = 1, \dots, p$, with probability at least $1 - \epsilon_{n,p} - o(p^{1-\delta})$, we have

$$\|\mathbf{A}\|_{\ell_{2,\infty}} = o(1) \quad \text{and} \quad \lim_{n \rightarrow \infty} \tau_j n^{1/2} = \theta_{j,j}^{-1/2}.$$

The first part of Theorem 1 presents that the error $\sqrt{n}(\hat{\mathbf{B}} - \mathbf{B})$ can be decomposed into a Gaussian term \mathbf{A} with zero mean and a bias term \mathbf{A} . The second part of this theorem shows that the bias term \mathbf{A} is asymptotically negligible with high probability. In particular, we prove that $\hat{\theta}_{j,j}^{1/2}$ converges to $\theta_{j,j}^{-1/2}$ with asymptotic probability one, which ensures the effectiveness of the inference in terms of the length of the confidence interval. It is worth noting that the product of the noise term τ_j and $n^{1/2}$ converges to the same constant as that of the estimator in Ref. [23] with asymptotic probability one. Since the noise factor is proportional to the variance of the estimator, the lengths of the confidence intervals for the two estimators are theoretically equal. Based on the conclusion in this theorem, we immediately obtain the asymptotic properties for all elements of the proposed two-step projection estimator, as shown in the following corollary.

Corollary 1. Under the conditions in Theorem 1, with a given significance level α , we further have

$$\lim_{n \rightarrow \infty} \mathbb{P} \left\{ \frac{\sqrt{n} |\hat{b}_{j,k} - b_{j,k}|}{\hat{\theta}_{j,j}^{1/2} \sigma} \leq \Phi^{-1} \left(1 - \frac{\alpha}{2} \right) \right\} = 1 - \alpha$$

for any $j = 1, \dots, p$ and $k = 1, \dots, q$.

Note that Corollary 1 still holds if the noise level σ is replaced by a consistent estimator $\hat{\sigma}$. Therefore this corollary provides theoretical guarantees for the constructed confidence interval in (6).

4 Simulation studies

In this section, we conduct simulation studies to investigate the performance of the proposed method compared with the generalization of LDPE[22] for multi-task regression[23] (denoted by MLDPE for simplicity). The implementation of $\hat{\mathbf{B}}_{CV}$ with $\ell_{2,1}$ group regularization is performed via the R package RMTL[29]. Based on $\hat{\mathbf{B}}_{CV}$, we obtain a consistent estimator $\hat{\sigma}$ for σ by applying the method stated in Section 2.2. Moreover, DC-SIS[25] is utilized to obtain a suitable prescreened set \hat{S} for those important predictors. To accurately control the size of \hat{S} , we take the least squares estimates on the subsets of \hat{S} and then use a BIC-type criterion[30] to choose the best subset.

In terms of the generation methods of the design matrix and error matrix, we conduct some simulations based on the following two models in both the sparse setting and the approximately sparse setting.

Model 1: The rows of the design matrix \mathbf{X} are sampled as independent and identically distributed copies from $N(\mathbf{0}, \Sigma_X)$, where $\Sigma_X = (0.5^{|i-j|})_{p \times p}$. The error items of \mathbf{E} are independent and identically distributed $N(0, \sigma^2)$ with $\sigma = 1$.

Model 2: The entries of the design matrix $\mathbf{X} = (x_{i,j})$ are Bernoulli random variables with a success probability of 0.8. All the columns of \mathbf{X} are centered to have zero mean. The entries of \mathbf{E} are generated from a t -distribution with 10 degrees of freedom.

We set the number of responses $q = 200$. Then, the sample size n , the number of predictors p and the coefficient matrices \mathbf{B} in different settings are constructed as follows:

Sparse setting: We set $(n, p) = (100, 200), (150, 400), (200, 800)$, respectively. Similar to Ref. [31], the elements in the first five rows of \mathbf{B} are drawn from a uniform distribution on $[-5, -1] \cup [1, 5]$, and the elements in other rows are set to be 0.

Approximately sparse setting: We set $(n, p) = (100, 400), (150, 600), (200, 1000)$, respectively. Similar to Refs. [22, 24], the j th important signal satisfies $\|\mathbf{b}^j\| = 3\lambda_{\text{univ}}$ with $\lambda_{\text{univ}} = \sqrt{2 \log(p)/n}$ for $j = 40, 80, 120, 160, 200$, and $\|\mathbf{b}^j\| = 3\lambda_{\text{univ}}/j^2$ for all other j . More specifically, to generate the j th row of \mathbf{B} with $\|\mathbf{b}^j\| = 3\lambda_{\text{univ}}$, we first generate a q -dimensional vector $\mathbf{v} = (v_1, \dots, v_q)$ with items $v_k \sim U[0, 1]$, $k = 1, \dots, q$. Then, we normalize \mathbf{v} such that $\|\mathbf{v}\| = 1$. Finally, we set $\mathbf{b}^j = 3\lambda_{\text{univ}} \mathbf{v}$.

The primary purpose of our simulation is to yield the 95% confidence intervals for the regression coefficients \mathbf{B} . In each setting, we run 100 replications and calculate the same three performance measures as those in Ref. [24]: the average coverage probability for all regression coefficients (CPA), the average coverage probability for important coefficients (CPI), and the average length of confidence intervals for all regression coefficients (Length).

Tables 1 and 2 summarize the results for the two methods

Table 1. Comparison of performance measures for two methods in the sparse setting.

Model		Method	TPE	MLDPE	TPE	MLDPE	
Case 1: $n = 100, p = 200, q = 200$		Model 1	$\hat{\sigma} = 1.006$		$\hat{\sigma} = 1.000$		
			CPA	0.9508 (0.0064)	0.9531 (0.0041)	0.9498 (0.0016)	0.9512 (0.0019)
			CPI	0.9474 (0.0104)	0.8516 (0.0331)	0.9471 (0.0088)	0.8480 (0.0212)
		Length	0.4323 (0.0114)	0.4200 (0.0111)	0.4297 (0.0021)	0.4172 (0.0029)	
		Model 2	$\hat{\sigma} = 1.256$		$\hat{\sigma} = 1.000$		
			CPA	0.9711 (0.0058)	0.9742 (0.0047)	0.9191 (0.0094)	0.9254 (0.0086)
CPI	0.9699 (0.0088)		0.9584 (0.0131)	0.9180 (0.0133)	0.8948 (0.0274)		
Length	0.5059 (0.0306)	0.4926 (0.0296)	0.4027 (0.0000)	0.3921 (0.0000)			
Case 2: $n = 150, p = 400, q = 200$		Model 1	$\hat{\sigma} = 1.024$		$\hat{\sigma} = 1.000$		
			CPA	0.9550 (0.0010)	0.9551 (0.0013)	0.9498 (0.0010)	0.9501 (0.0000)
			CPI	0.9570 (0.0068)	0.8732 (0.0267)	0.9519 (0.0075)	0.8647 (0.0259)
		Length	0.3535 (0.0028)	0.3534 (0.0028)	0.3455 (0.0000)	0.3455 (0.0000)	
		Model 2	$\hat{\sigma} = 1.285$		$\hat{\sigma} = 1.000$		
			CPA	0.9749 (0.0030)	0.9755 (0.0029)	0.9191 (0.0056)	0.9202 (0.0051)
CPI	0.9735 (0.0085)		0.9678 (0.0092)	0.9115 (0.0122)	0.9035 (0.0144)		
Length	0.4112 (0.0108)	0.4114 (0.0139)	0.3198 (0.0000)	0.3201 (0.0000)			
Case 3: $n = 200, p = 800, q = 200$		Model 1	$\hat{\sigma} = 1.039$		$\hat{\sigma} = 1.000$		
			CPA	0.9581 (0.0019)	0.9563 (0.0019)	0.9499 (0.0014)	0.9478 (0.0012)
			CPI	0.9606 (0.0076)	0.8863 (0.0090)	0.9536 (0.0069)	0.8736 (0.0112)
		Length	0.3117 (0.0010)	0.3124 (0.0036)	0.2997 (0.0000)	0.3005 (0.0000)	
		Model 2	$\hat{\sigma} = 1.279$		$\hat{\sigma} = 1.000$		
			CPA	0.9752 (0.0029)	0.9749 (0.0026)	0.9213 (0.0051)	0.9207 (0.0057)
CPI	0.9769 (0.0049)		0.974 (0.0044)	0.9245 (0.0076)	0.9158 (0.0104)		
Length	0.3544 (0.0132)	0.3546 (0.0130)	0.2760 (0.0000)	0.2762 (0.0000)			

in different settings. Clearly, the results in the sparse setting and approximately sparse setting are similar. In Gaussian settings, it can be seen from CPA (or CPI) that the average coverage probabilities for all regression coefficients (or important coefficients) of the proposed method are approximately 95%, while the average coverage probabilities for all regression coefficients (or important coefficients) of MLDPE deviate from 95% slightly. In view of Length, the average lengths of the confidence intervals for the two methods are roughly the same. In non-Gaussian settings, the performance of both TPE and MLDPE tends to worsen.

In each setting, we further set $\hat{\sigma} = 1$ to calculate performance measures for comparison. We can see that the performance of the proposed method remains stable, while the performance of MLDPE fluctuates slightly. In summary, the proposed method outperforms MLDPE in both the sparse setting and the approximately sparse setting.

5 Application to TCGA-OV data

The Cancer Genome Atlas (TCGA) is a cancer genomics program incorporating clinical data on human cancers and tumor subtypes, including aberrations in gene expression, epigenetics (miRNAs, methylation), and protein expression. In this section, we apply our methodology to the ovarian serous cyst-

adenocarcinoma (TCGA-OV) data downloaded from the TCGA website^①. We regarded the miRNA expression quantification obtained by the miRNA-seq experimental strategy as predictors, and protein expression quantification using the reversed-phase protein array technique as responses. Our goal is to explore the regulatory effect of miRNA on protein expression in ovarian cancer.

After preliminary data processing, the number of miRNAs was reduced to 1530, and the number of proteins was 216. Then, we utilize the DC-SIS method to perform feature screening on the predictors, select the first 400 important miRNAs as predictors, and choose the first 50 proteins with larger variance as response variables. Finally, we obtained $n = 300$ common samples with $p = 400$ miRNAs as predictors and $q = 50$ proteins as responses. It is worth noting that both predictors and responses are centered and normalized to have zero mean and a common ℓ_2 -norm \sqrt{n} .

By applying the proposed method, we obtain 95% confidence intervals for all unknown coefficients. As a result, the top 45 important miRNAs are selected in Table 3, 32 of which highlighted in bold are also chosen by MLDPE. These selected miRNAs play important regulatory roles in cancer. For example, compared to normal ovaries, some miRNAs are

① <https://portal.gdc.cancer.gov/>.

Table 2. Comparison of performance measures for two methods in the nonsparse setting.

	Model	Method	TPE	MLDPE	TPE	MLDPE		
Case 1: $n = 100, p = 400, q = 200$	Model 1		$\hat{\sigma} = 1.118$		$\hat{\sigma} = 1.000$			
			CPA	0.9683 (0.0095)	0.9686 (0.0103)	0.9474 (0.0013)	0.9480 (0.0013)	
			CPI	0.9677 (0.0102)	0.9671 (0.0122)	0.9443 (0.0059)	0.9458 (0.0054)	
		Length	0.4664 (0.0323)	0.4616 (0.0318)	0.4172 (0.0012)	0.4129 (0.0011)		
		Model 2		$\hat{\sigma} = 1.290$		$\hat{\sigma} = 1.000$		
				CPA	0.9760 (0.0048)	0.9756 (0.0050)	0.9208 (0.0097)	0.9202 (0.0094)
	CPI			0.9754 (0.0050)	0.9759 (0.0048)	0.9250 (0.0108)	0.9239 (0.0112)	
	Length	0.5109 (0.0235)	0.5058 (0.0232)	0.3961 (0.0000)	0.3921 (0.0000)			
	Case 2: $n = 150, p = 600, q = 200$	Model 1		$\hat{\sigma} = 1.049$		$\hat{\sigma} = 1.000$		
				CPA	0.9578 (0.0022)	0.9577 (0.0024)	0.9484 (0.0012)	0.9481 (0.0013)
				CPI	0.9573 (0.0057)	0.9558 (0.0073)	0.9502 (0.0055)	0.9463 (0.0048)
			Length	0.3581 (0.0053)	0.3583 (0.0052)	0.3428 (0.0000)	0.3429 (0.0000)	
Model 2				$\hat{\sigma} = 1.307$		$\hat{\sigma} = 1.000$		
				CPA	0.9767 (0.0063)	0.9765 (0.0065)	0.9198 (0.0055)	0.9197 (0.0055)
		CPI		0.9746 (0.0072)	0.9740 (0.0073)	0.9162 (0.0102)	0.9173 (0.0102)	
Length		0.4184 (0.0335)	0.4183 (0.0332)	0.3203 (0.0000)	0.3201 (0.0000)			
Case 3: $n = 200, p = 1000, q = 200$		Model 1		$\hat{\sigma} = 1.045$		$\hat{\sigma} = 1.000$		
				CPA	0.9580 (0.0024)	0.9582 (0.0023)	0.9485 (0.0000)	0.9487 (0.0000)
				CPI	0.9581 (0.0073)	0.9582 (0.0077)	0.9500 (0.0066)	0.9502 (0.0066)
			Length	0.3132 (0.0041)	0.3132 (0.0041)	0.2997 (0.0000)	0.2998 (0.0000)	
	Model 2			$\hat{\sigma} = 1.345$		$\hat{\sigma} = 1.000$		
				CPA	0.9783 (0.0035)	0.9781 (0.0035)	0.9132 (0.0051)	0.9127 (0.0051)
		CPI		0.9774 (0.0050)	0.9768 (0.0049)	0.9093 (0.0110)	0.9102 (0.0080)	
	Length	0.3726 (0.0159)	0.3727 (0.0158)	0.2771 (0.0000)	0.2772 (0.0000)			

Table 3. 45 miRNAs selected by TPE. The miRNAs also selected by MLDPE are highlighted in bold.

hsa-mir-486-2	hsa-mir-181a-2	hsa-mir-16-1	hsa-mir-24-1	hsa-mir-769
hsa-mir-26a-1	hsa-mir-214	hsa-mir-16-2	hsa-mir-130a	hsa-mir-130b
hsa-mir-125b-1	hsa-mir-99a	hsa-mir-9-1	hsa-mir-22	hsa-mir-377
hsa-mir-194-2	hsa-mir-101-2	hsa-mir-508	hsa-mir-21	hsa-mir-1247
hsa-mir-26a-2	hsa-mir-200a	hsa-mir-605	hsa-mir-30c-2	hsa-mir-132
hsa-mir-199a-1	hsa-mir-24-2	hsa-mir-766	hsa-mir-29a	hsa-mir-433
hsa-mir-486-1	hsa-mir-365b	hsa-mir-378c	hsa-mir-150	hsa-let-7f-2
hsa-mir-30c-1	hsa-mir-181a-1	hsa-mir-654	hsa-mir-223	hsa-mir-182
hsa-mir-509-3	hsa-mir-127	hsa-mir-378a	hsa-mir-432	hsa-mir-493

dysregulated in ovarian cancer. It is worth noting that the importance is reflected by the magnitude of the sum of absolute values for these estimated coefficients over all 50 proteins. Among these selected miRNAs, hsa-mir-182, hsa-mir-200a, hsa-mir-223, and hsa-mir-16 are upregulated, and hsa-mir-432, hsa-mir-493, hsa-mir-9 and hsa-mir-377 are downregulated^[32]. Due to the dysregulation of these miRNAs, some of

them could potentially be used as diagnostic biomarkers, including hsa-mir-214 and hsa-mir-21^[33].

Experimental results show that the average length of confidence intervals obtained by TPE method is 0.4372, while MLDPE method yields the average length of 0.4136. It is clear that the average lengths of both methods are around the same level. The No.1 miRNA chosen by both TPE and

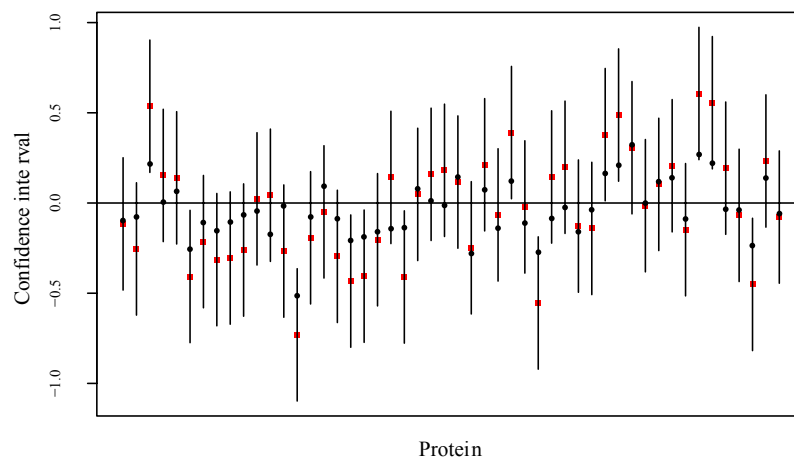


Fig. 1. Estimates of the unknown coefficients of miRNA hsa-mir-486-2 (red squares for TPE and black dots for MLDPE) and the corresponding 95% confidence intervals (obtained by TPE) over all 50 proteins.

MLDPE is hsa-mir-486-2, which was also identified as a potential biomarker for lung adenocarcinoma^[34]. For the unknown coefficients of hsa-mir-486-2, Fig. 1 displays their estimates and the corresponding 95% confidence intervals over all 50 proteins, which further justifies the important regulatory roles of hsa-mir-486-2.

6 Conclusions

In this paper, we develop a new inference methodology based on the two-step projection estimator (TPE) in high-dimensional multi-task regression. The proposed estimator is established in a row-wise manner, which has the benefit of reducing the estimation bias induced by these important signals. In addition, we provide strict theoretical guarantees for our method, including asymptotic normality and corresponding confidence intervals. Moreover, the numerical results of the proposed method indicate that the proposed method works quite well. Specifically, we apply this approach to an ovarian cancer dataset and identify several miRNAs that are closely associated with protein expression levels. Results demonstrate that these miRNAs can potentially serve as biomarkers in disease research, aiding in the diagnosis of the ovarian cancer.

It would be interesting to extend our method to more general settings, such as multi-response linear regression models with measurement errors and generalized linear models. It is also of interest to build a relationship between the two-step projection technique and the partially penalized procedure in high-dimensional multi-response settings. These generalizations are interesting topics for future research.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (12101584), the China Postdoctoral Science Foundation (2021TQ0326, 2021M703100), Fundamental Research Funds for the Central Universities (WK2040000047), Hefei Postdoctoral Research Project Funds in 2021, and Anhui Postdoctoral Research Project Funds in 2021.

Conflict of interest

The authors declare that they have no conflict of interest.

Biographies

Yuanli Ma is currently a master student at the University of Science and Technology of China. Her research mainly focuses on big data problems.

Yang Li is currently a postdoctoral researcher at the University of Science and Technology of China (USTC). He received his Ph.D. degree in Statistics from USTC in 2021. His research interests include high-dimensional statistical inference and distributed learning.

References

- [1] Lounici K, Pontil M, Tsybakov A B, et al. Taking advantage of sparsity in multi-task learning. arXiv:0903.1468, **2009**.
- [2] Obozinski G, Taskar B, Jordan M I. Joint covariate selection and joint subspace selection for multiple classification problems. *Stat. Comput.*, **2010**, *20* (2): 231–252.
- [3] Lounici K, Pontil M, Van De Geer S, et al. Oracle inequalities and optimal inference under group sparsity. *Ann. Statist.*, **2011**, *39* (4): 2164–2204.
- [4] Wang H, Nie F, Huang H, et al. Identifying quantitative trait loci via group-sparse multitask regression and feature selection: An imaging genetics study of the ADNI cohort. *Bioinformatics*, **2012**, *28* (2): 229–237.
- [5] Greenlaw K, Szefer E, Graham J, et al. A Bayesian group sparse multi-task regression model for imaging genetics. *Bioinformatics*, **2017**, *33* (16): 2513–2522.
- [6] Zhou J J, Cho M H, Lange C, et al. Integrating multiple correlated phenotypes for genetic association analysis by maximizing heritability. *Human Heredity*, **2015**, *79* (2): 93–104.
- [7] Kim S, Sohn K-A, Xing E P. A multivariate regression approach to association analysis of a quantitative trait network. *Bioinformatics*, **2009**, *25* (12): i204–i212.
- [8] Mørk S, Pletscher-Frankild S, Palleja Caro A, et al. Protein-driven inference of miRNA-disease associations. *Bioinformatics*, **2014**, *30* (3): 392–397.
- [9] Gommans W M, Berezikov E. Controlling miRNA regulation in disease. In: Next-Generation MicroRNA Expression Profiling Technology: Methods and Protocols. Totowa, NJ: Humana Press, **2012**: 1–18.
- [10] Izenman A J. Reduced-rank regression for the multivariate linear

- model. *J. Multivariate Anal.*, **1975**, 5 (2): 248–264.
- [11] Velu R, Reinsel G C. *Multivariate Reduced-Rank Regression: Theory and Applications*. New York: Springer Science & Business Media, **1998**.
- [12] Anderson T W. Asymptotic distribution of the reduced rank regression estimator under general conditions. *Ann. Statist.*, **1999**, 27 (4): 1141–1154.
- [13] Uematsu Y, Fan Y, Chen K, et al. SOFAR: Large-scale association network learning. *IEEE Trans. Inform. Theory*, **2019**, 65 (8): 4924–4939.
- [14] Zheng Z, Li Y, Wu J, et al. Sequential scaled sparse factor regression. *J. Bus. Econom. Statist.*, **2022**, 40 (2): 595–604.
- [15] Yuan M, Ekici A, Lu Z, et al. Dimension reduction and coefficient estimation in multivariate linear regression. *The Journal of the Royal Statistical Society, Series B: Statistical Methodology*, **2007**, 69 (3): 329–346.
- [16] Bunea F, She Y, Wegkamp M H. Joint variable and rank selection for parsimonious estimation of high-dimensional matrices. *Ann. Statist.*, **2012**, 40 (5): 2359–2388.
- [17] Chen L, Huang J Z. Sparse reduced-rank regression for simultaneous dimension reduction and variable selection. *J. Amer. Statist. Assoc.*, **2012**, 107 (500): 1533–1545.
- [18] Chen K, Chan K-S, Stenseth N C. Reduced rank stochastic regression with a sparse singular value decomposition. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, **2012**, 74 (2): 203–221.
- [19] Obozinski G, Wainwright M J, Jordan M I. Support union recovery in high-dimensional multivariate regression. *Ann. Statist.*, **2011**, 39 (1): 1–47.
- [20] Turlach B A, Venables W N, Wright S J. Simultaneous variable selection. *Technometrics*, **2005**, 47 (3): 349–363.
- [21] Quattoni A, Carreras X, Collins M, et al. An efficient projection for $\ell_{1,\infty}$ regularization. In: *Proceedings of the 26th Annual International Conference on Machine Learning*. New York: ACM, **2009**: 857–864.
- [22] Zhang C-H, Zhang S S. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **2014**, 76 (1): 217–242.
- [23] Chevalier J-A, Salmon J, Gramfort A, et al. Statistical control for spatio-temporal MEG/EEG source imaging with desparsified multi-task lasso. In: *Advances in Neural Information Processing Systems 33*. Red Hook, NY: Curran Associates, Inc., **2020**: 1759–1770.
- [24] Li Y, Zheng Z, Zhou J, et al. High-dimensional inference via hybrid orthogonalization. arXiv:2111.13391, **2012**.
- [25] Li R, Zhong W, Zhu L. Feature screening via distance correlation learning. *Journal of the American Statistical Association*, **2012**, 107 (499): 1129–1139.
- [26] Székely G J, Rizzo M L, Bakirov N K. Measuring and testing dependence by correlation of distances. *Ann. Statist.*, **2007**, 35 (6): 2769–2794.
- [27] Reid S, Tibshirani R, Friedman J. A study of error variance estimation in lasso regression. *Statist. Sinica*, **2016**, 26: 35–67.
- [28] Ye F, Zhang C H. Rate minimaxity of the lasso and Dantzig selector for the ℓ_q loss in ℓ_r balls. *Journal of Machine Learning Research*, **2010**, 11: 3519–3540.
- [29] Cao H, Zhou J, Schwarz E. RMTL: an R library for multi-task learning. *Bioinformatics*, **2019**, 35 (10): 1797–1798.
- [30] Sakurai T, Fujikoshi Y. High-dimensional properties of information criteria and their efficient criteria for multivariate linear regression models with covariance structures. **2017**. <http://www.math.sci.hiroshima-u.ac.jp/stat/TR/TR17/TR17-13.pdf>. Accessed August 1, 2022.
- [31] Li Y, Nan B, Zhu J. Multivariate sparse group lasso for the multivariate multiple linear regression with an arbitrary group structure. *Biometrics*, **2015**, 71 (2): 354–363.
- [32] Aziz N B, Mahmudunnabi R G, Umer M, et al. MicroRNAs in ovarian cancer and recent advances in the development of microRNA-based biosensors. *Analyst*, **2020**, 145 (6): 2038–2057.
- [33] Wu Y D, Li Q, Zhang R S, et al. Circulating microRNAs: Biomarkers of disease. *Clinica Chimica Acta*, **2021**, 516: 46–54.
- [34] Ren Z P, Hou X B, Tian X D, et al. Identification of nine microRNAs as potential biomarkers for lung adenocarcinoma. *FEBS Open Bio*, **2019**, 9 (2): 315–327.

Appendix

A.1 Proof of Proposition 1

The proof of Proposition 1 is similar to that of Theorem 1 in Ref. [25]. We first verify whether Conditions (C1) and (C2) required by Theorem 1 of Ref. [25] still hold in this paper. Since both \mathbf{x} and \mathbf{e} obey the multivariate normal distribution, we can deduce that \mathbf{x} and \mathbf{y} have a multivariate normal distribution because \mathbf{x} is independent of \mathbf{e} . Thus, Condition (C1) in Ref. [25] still holds in the framework of this paper. Recall that S_A is defined as follows to contain the indices of all important predictors:

$$S_A = \{i \in \{1, \dots, p\} : \omega_i \geq cn^{-\kappa}\},$$

where $c > 0$ and $0 \leq \kappa < 1/2$ are some constants. As a result, Condition (C2) in Ref. [25] also holds in this paper.

Based on the above facts, applying the second conclusion of Theorem 1 in Ref. [25] yields

$$\mathbb{P}\{S_A \subset \hat{S}\} \geq 1 - O(s\{\exp\{-c_1 n^{1-2(\kappa+\gamma)}\} + n \exp\{-c_2 n^\gamma\}\})$$

for any $0 < \gamma < 1/2 - \kappa$, where c_1 and c_2 are some positive constants. Similar to Ref. [25], by choosing the optimal order $\gamma = (1 - 2\kappa)/3$, we further have $\mathbb{P}\{S_A \subset \hat{S}\} \geq 1 - O(s\exp\{-c_0 n^{(1-2\kappa)/3}\})$ with constant $c_0 > 0$, which completes the proof of Proposition 1.

A.2 Proof of Theorem 1

Under the conditions in Theorem 1, we can deduce that the theoretical properties of \mathbf{z}_j are the same as those in Ref. [24]. Therefore, the details for deriving the properties of \mathbf{z}_j are omitted here. We proceed to prove the main results in Theorem 1. For any $j \in \{1, \dots, p\}$, in view of the definition $\hat{\mathbf{b}}^j = \frac{\mathbf{z}_j^T \mathbf{Y}}{\mathbf{z}_j^T \mathbf{x}_j}$, plugging model (2) to (3) yields

$$\sqrt{n}(\hat{\mathbf{b}}^j - \mathbf{b}^j) = \sqrt{n} \left(\frac{\mathbf{z}_j^\top (\sum_k \mathbf{x}_k \mathbf{b}^k + \mathbf{E})}{\mathbf{z}_j^\top \mathbf{x}_j} - \mathbf{b}^j \right) = \sqrt{n} \sum_k \frac{\mathbf{z}_j^\top \mathbf{x}_k}{\mathbf{z}_j^\top \mathbf{x}_j} \mathbf{b}^k + \sqrt{n} \frac{\mathbf{z}_j^\top \mathbf{E}}{\mathbf{z}_j^\top \mathbf{x}_j} - \sqrt{n} \mathbf{b}^j.$$

With the aid of $\mathbf{z}_j^\top \mathbf{x}_k = 0$ for $k \in \hat{S} \setminus \{j\}$, some simple algebra shows

$$\sqrt{n}(\hat{\mathbf{b}}^j - \mathbf{b}^j) = \sqrt{n} \frac{\mathbf{z}_j^\top \mathbf{E}}{\mathbf{z}_j^\top \mathbf{x}_j} + \sqrt{n} \sum_{k \neq j} \frac{\mathbf{z}_j^\top \mathbf{x}_k}{\mathbf{z}_j^\top \mathbf{x}_j} \mathbf{b}^k = \sqrt{n} \frac{\mathbf{z}_j^\top \mathbf{E}}{\mathbf{z}_j^\top \mathbf{x}_j} + \sqrt{n} \sum_{k \in \hat{S}^c, k \neq j} \frac{\mathbf{z}_j^\top \mathbf{x}_k}{\mathbf{z}_j^\top \mathbf{x}_j} \mathbf{b}^k := \mathbf{A}_{j \cdot} + \mathbf{A}_{j \cdot}^c,$$

where $\mathbf{A}_{j \cdot}$ and $\mathbf{A}_{j \cdot}^c$ are the j th row of \mathbf{A} and \mathbf{A}^c , respectively. In what follows, we introduce the properties of \mathbf{A} and \mathbf{A}^c in the following two parts.

Part 1: In this part, we will prove that $\mathbf{A}_{j \cdot} = \sqrt{n} \frac{\mathbf{z}_j^\top \mathbf{E}}{\mathbf{z}_j^\top \mathbf{x}_j} \sim N(\mathbf{0}, \widehat{\theta}_{jj} \sigma^2 \mathbf{I})$, or $\mathbf{E}^\top \mathbf{z}_j \sim N(\mathbf{0}, \|\mathbf{z}_j\|^2 \sigma^2 \mathbf{I}_{q \times q})$, equivalently. The proof is similar to that of Proposition 2.1 in Ref. [23]. Clearly, $\mathbf{E}^\top \mathbf{z}_j$ is a Gaussian vector and its mean is $\mathbf{0}$, we then only need to prove that the covariance of $\mathbf{E}^\top \mathbf{z}_j$ equals $\|\mathbf{z}_j\|^2 \sigma^2 \mathbf{I}_{q \times q}$. By the definition of covariance, it follows that

$$\text{cov}(\mathbf{E}^\top \mathbf{z}_j) = \mathbb{E}(\mathbf{E}^\top \mathbf{z}_j \mathbf{z}_j^\top \mathbf{E}) - \mathbb{E}(\mathbf{E}^\top \mathbf{z}_j) \mathbb{E}(\mathbf{z}_j^\top \mathbf{E}) = \mathbb{E}(\mathbf{E}^\top \mathbf{z}_j \mathbf{z}_j^\top \mathbf{E}) := [\mathbf{M}^{(j)}]_{q \times q}.$$

Then, for any $1 \leq t, t' \leq q$, some simple algebra gives

$$\mathbf{M}_{t,t'}^{(j)} = \mathbb{E}(\mathbf{u}_i^\top \mathbf{E}^\top \mathbf{z}_j \mathbf{z}_j^\top \mathbf{E} \mathbf{u}_i) = \mathbb{E}(\mathbf{e}_i^\top \mathbf{z}_j \mathbf{z}_j^\top \mathbf{e}_i) = \mathbb{E}(\mathbf{z}_j^\top \mathbf{e}_i \mathbf{e}_i^\top \mathbf{z}_j) = z_{j,1} \mathbb{E}(e_{1,t'} e_{1,t}) z_{j,1} + \dots + z_{j,n} \mathbb{E}(e_{n,t'} e_{n,t}) z_{j,n}.$$

where $\mathbf{z}_j = (z_{j,1}, \dots, z_{j,n})^\top$, and $\mathbf{u}_i = (0, \dots, 0, \underbrace{1}_{i\text{th}}, 0, \dots, 0)^\top$ is a unit column vector. Since $\mathbf{e}^i \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_{q \times q})$, we can derive that

$$\text{cov}(e_{i,t} e_{i,t'}) = \begin{cases} \sigma^2, & t = t'; \\ 0, & t \neq t'. \end{cases}$$

Thus, if $t = t'$, combining the above results shows

$$\mathbf{M}_{t,t}^{(j)} = \sum_{i=1}^n z_{j,i}^2 \sigma^2 = \|\mathbf{z}_j\|^2 \sigma^2,$$

which further entails $\mathbf{A}_{j \cdot} \sim N(\mathbf{0}, \widehat{\theta}_{jj} \sigma^2 \mathbf{I})$. It completes the proof of this part.

Part 2: In this part, we will prove that $\|\mathbf{A}^c\|_{\ell_{2,\infty}} = o(1)$ with asymptotic probability one. For any $1 \leq j \leq p$, it follows that

$$\|\mathbf{A}_{j \cdot}^c\| = \sqrt{n} \left\| \sum_{k \in \hat{S}^c, k \neq j} \frac{\mathbf{z}_j^\top \mathbf{x}_k}{\mathbf{z}_j^\top \mathbf{x}_j} \mathbf{b}^k \right\| \leq \sqrt{n} \left(\max_{k \in \hat{S}^c, k \neq j} \left| \frac{\mathbf{z}_j^\top \mathbf{x}_k}{\mathbf{z}_j^\top \mathbf{x}_j} \right| \right) \sum_{k \in \hat{S}^c} \|\mathbf{b}^k\|.$$

In view of the properties of \mathbf{z}_j in Ref. [24], for sufficiently large n and $\delta \geq 1$, it follows that

$$\mathbb{P} \left\{ \max_{k \in \hat{S}^c, k \neq j} \left| \frac{\mathbf{z}_j^\top \mathbf{x}_k}{\mathbf{z}_j^\top \mathbf{x}_j} \right| \leq C \sqrt{\log(p)/n} \right\} \geq 1 - o(p^{1-\delta}),$$

where C is a positive constant.

Denote by events

$$\mathcal{E} = \{S_A \subset \hat{S}\} \quad \text{and} \quad \mathcal{E}_* = \left\{ \max_{k \in \hat{S}^c, k \neq j} \left| \frac{\mathbf{z}_j^\top \mathbf{x}_k}{\mathbf{z}_j^\top \mathbf{x}_j} \right| \leq C \sqrt{\log(p)/n} \right\}.$$

Our theoretical analysis will depend on the event $\mathcal{E}_1 = \mathcal{E} \cap \mathcal{E}_*$. It follows from the definition of the suitable prescreened set \hat{S} that

$$\mathbb{P}(\mathcal{E}_1) = 1 - \mathbb{P}(\mathcal{E}^c \cup \mathcal{E}_*^c) \geq 1 - \mathbb{P}(\mathcal{E}^c) - \mathbb{P}(\mathcal{E}_*^c) \geq 1 - \epsilon_{n,p} - o(p^{1-\delta}).$$

In view of $\sum_{k \in \hat{S}^c} \|\mathbf{b}^k\| = o(1/\sqrt{\log(p)})$ in the second part of Condition 3, combining the above results yields

$$\|\mathbf{A}^c\|_{\ell_{2,\infty}} = \max_{1 \leq j \leq p} \|\mathbf{A}_{j \cdot}^c\| \leq \sqrt{n} \cdot C \sqrt{\log(p)/n} \cdot o(1/\sqrt{\log(p)}) = o(1)$$

with probability at least $1 - \epsilon_{n,p} - o(p^{1-\delta})$.

In addition, the convergence of $\tau_j n^{1/2}$ is essentially the same as that of Ref. [24]. Therefore, it is omitted here. Finally, it completes the proof of Theorem 1.