# Estimation of Low Rank High Dimensional Multivariate Linear Models for Multi-response Data *

Changliang Zou
School of Statistics and Data Sciences
Nankai University

Yuan Ke
Department of Statistics
University of Georgia

Wenyang Zhang †
Department of Mathematics
The University of York, UK

January 26, 2021

**Abstract**

In this paper, we study low rank high dimensional multivariate linear models (LRMLM) for high dimensional multi-response data. We propose an intuitively appealing estimation approach, and develop an algorithm for implementation purposes. Asymptotic properties are established in order to justify the estimation procedure theoretically. Intensive simulation studies are also conducted to demonstrate performance when the sample size is finite, and a comparison is made with some popular methods from the literature. The results show the proposed estimator outperforms all of the alternative methods under various circumstances. Finally, using our suggested estimation procedure we apply the LRMLM to analyse an environmental data set and predict concentrations of PM2.5 at the locations

*This research is supported by National Natural Science Foundation of China (Grant Numbers 11931014, 11931001, 11690015, 11925106) and NSF of Tianjin Grant 18JCJQJC46000.

†The corresponding author, Department of Mathematics, University of York, York, YO10 5DD, United Kingdom, Email: wenyang.zhang@york.ac.uk.

concerned. The results illustrate how the proposed method provides more accurate predictions than the alternative approaches.

**KEY WORDS**: BIC, cross-validation, high dimensionality, low rank, multivariate linear models, penalised least squares estimation.

**SHORT TITLE**: LRMLM.

# 1 Introduction

It is common to find multi-response data in many real life problems. Component-wise analysis is clearly not a good choice for multi-response data analysis, because it does not fully make use of the information available. For example, the observations of other components may contain the information for the component of interest, and such information would be completely overlooked by component-wise analysis, therefore, the resulting estimators would not be as efficient as we can expect. It is necessary to take multivariate analysis approach for multi-response data analysis. The most commonly used multivariate regression models are the multivariate linear models. The research in the multivariate linear models can be at least traced back to Anderson (1951). There is much literature after Anderson (1951) about the classic multivariate linear models, see the references in Reinsel and Velu (1998) and Anderson (2004).

With the surge in high dimensional data analysis in the past more than a decade, the multivariate linear models in high dimensional setting are attracting more and more attention than ever before. Many interesting developments in low rank high dimensional multivariate linear models have appeared in literature, see Yuan et al. (2007), Negahban and Wainwright (2011), Obozinski et al. (2011), Kong et al. (2017), Bing and Wegkamp (2019), Raskutti et al. (2019), Zheng et al. (2019) and the references therein.

A commonly used approach to deal with the low rank coefficient matrix in a multivariate linear model is based on the idea of decomposing the coefficient matrix, say $\mathbf{A}$ with rank $r$ and size $p \times q$, to $\mathbf{CDQ}$, where $\mathbf{C}$ and $\mathbf{Q}$ are two matrices of size $p \times \hat{r}$ and $\hat{r} \times q$, respectively, and $\mathbf{D}$ is a diagonal matrix of

size $\hat{r}$, where $\hat{r}$ is an estimator of $r$. The estimation of $r$ plays a key role for the success of the approach used. Different approaches may end up with different ways to estimate $r$, see Yuan et al. (2007) and Bing and Wegkamp (2019). Although the existing approaches for estimating $r$ enjoy nice asymptotic properties, when implementing them, we often come up against a dilemma: we create a new unknown parameter in order to estimate an unknown parameter, this is because we have to select a tuning parameter in the estimation of $r$. In addition to that, as far as the estimation of $\mathbf{A}$ is concerned, which is the ultimate goal for multivariate linear models, even if we knew the rank $r$, in order to get the estimator of $\mathbf{A}$ based on the decomposition, we would have to estimate $\mathbf{C}$, $\mathbf{Q}$ and $\mathbf{D}$. Even with the constraints coming with the decomposition, we may have to estimate at least $(p+q)r$ unknown parameters, which is more than the unknown parameters we need to estimate without using the decomposition when $r > pq/(p+q)$. That implies we may end up with a better estimator of $\mathbf{A}$ if we simply apply the standard least squares estimation for multivariate linear models when $r > pq/(p+q)$, which clearly shows the limitation of the decomposition based approach.

In this paper, we are going to propose an estimation procedure for the low rank multivariate linear models, in which we only need to estimate $r(p+q)-r^2$ unknown parameters in order to get the estimator of $\mathbf{A}$. We can easily show $r(p+q)-r^2 \leq pq$, because $r \leq \min(p,\ q)$. Intuitively speaking, the proposed estimation procedure would be more efficient than either of the standard least squares estimation and the decomposition based approach. This conclusion is confirmed to be true by both the asymptotic theory established in Section 4 and simulation studies in Section 5. As part of the proposed estimation procedure, the rank of $\mathbf{A}$ is estimated by the BIC, which is free of tuning parameter. We will show the resulting estimator enjoys good theoretical properties and performs well in simulation studies.

Another advantage of the proposed estimation procedure is it clearly appreciates the high dimensionality by directly imposing a penalty on those entries of $\mathbf{A}$ in question, which makes the proposed estimation procedure easily accom-

modate the high dimensional cases and enjoy the function of feature selection.

In the context of multi-response data analysis, the proposed estimation procedure also comes with a very nice practical implication, which is the impacts of explanatory variables on some responses are linear combinations of the impacts on certain responses when the matrix coefficient is of low rank, this would be very helpful when it comes to interpreting the results for a given real dataset, and may lead to some interesting findings in the discipline which the dataset comes from.

To implement the proposed estimation procedure, we have also developed an algorithm for the estimation. Our simulation studies show the proposed algorithm is fast and accurate.

The rest of this paper is organised as follows: we begin with a detailed description of the models we are going to address in Section 2. The proposed estimation procedure and associated computational algorithm are described in Section 3. The asymptotic properties of the estimators obtained by the proposed estimation procedure are presented in Section 4. Section 5 is devoted to simulation studies, in which we will examine how well the proposed estimation works. Finally, in Section 6, we apply the low rank multivariate linear models together with the proposed estimation procedure to analyse an environmental data set and predict the concentrations of PM2.5 at the locations concerned. The results show the proposed method provides more accurate prediction than other methods. We leave the theoretical proofs of all asymptotic properties in the Appendix.

## 2 The low rank high dimensional multivariate linear models

To give a generic description of the models we are going to address, we use $Y$ to denote the vector of all response variables, $X$ the vector of all covariates. Without any confusion, from now on, we call $Y$ the response variable, $X$ the covariate. We assume $Y$ is of $q$ dimension, $X$ is of $p$ dimension. $p$ and $q$

may tend to $\infty$ when sample size tends to $\infty$. The low rank high dimensional multivariate linear models which we are going to address in this paper are

$$Y = \mathbf{A}^\top X + \boldsymbol{\epsilon}, \tag{2.1}$$

where $\mathbf{A}$ is a $p \times q$ unknown matrix of unknown rank $r$, $r < q < p$, $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_q)^\top$ is a $q$ dimensional random error, and

$$\mathbb{E}(\boldsymbol{\epsilon} \mid X) = \mathbf{0}, \quad \text{cov}(\boldsymbol{\epsilon} \mid X) = \boldsymbol{\Sigma}.$$

Like Yuan et al. (2007), Bing and Wegkamp (2019) and Raskutti et al. (2019), we assume $\boldsymbol{\Sigma} = \sigma^2 \mathbf{I}_q$, and $\sigma^2$ is unknown.

Suppose we have a sample $(X_i^\top,\ Y_i^\top)$, $i = 1,\ \cdots,\ n$, from $(X^\top,\ Y^\top)$, the model for the sample can be written as

$$\mathbf{Y} = \mathbf{X}\mathbf{A} + \mathbf{E} \tag{2.2}$$

where $\mathbf{Y} = (Y_1,\ \cdots,\ Y_n)^\top$, $\mathbf{X} = (X_1,\ \cdots,\ X_n)^\top$, and $\mathbf{E} = (\boldsymbol{\epsilon}_1,\ \cdots,\ \boldsymbol{\epsilon}_n)^\top$.

## 3 Estimation procedure

Throughout this paper, for any matrix $\boldsymbol{\Omega} = (w_{ij})$ of size $p \times q$ and any vector $\mathbf{b} = (b_1,\ \cdots,\ b_p)$, we define

$$\|\boldsymbol{\Omega}\|_1 = \sum_{i=1}^{p}\sum_{j=1}^{q} |w_{ij}|, \quad \|\boldsymbol{\Omega}\| = \left(\sum_{i=1}^{p}\sum_{j=1}^{q} w_{ij}^2\right)^{1/2}, \|\mathbf{b}\|_2 = (\mathbf{b}^\top \mathbf{b})^{1/2}, \quad \|\mathbf{b}\|_1 = \sum_{i=1}^{p} |b_i|.$$

For any integers $1 \le j_1 < j_2 < \cdots < j_r \le q$, the complement set of set $\mathcal{D}_r$, $\mathcal{D}_r = \{j_1, \cdots,\ j_r\}$, is denoted by $\mathcal{D}_r^c$, that is $\mathcal{D}_r^c = \{1,\ 2,\ \cdots,\ q\} \setminus \mathcal{D}_r$. Let $\mathbf{Y}^i$ be the $i$th column of $\mathbf{Y}$,

$$\mathbf{Y}_{\mathcal{D}_r} = (\mathbf{Y}^{j_1},\ \cdots,\ \mathbf{Y}^{j_r}), \quad \mathbb{H}_k(r) = \{\mathcal{D}_k : \text{rank}(\mathbf{A}_{\mathcal{D}_k}) = r\},$$

$$\bar{\mathbb{H}}_k(r) = \{\mathcal{D}_k : \text{rank}(\mathbf{A}_{\mathcal{D}_k}) < r\}, \quad \mathbb{G}(k) = \{\mathbf{B} \in \mathbb{R}^{p\times k} : \text{rank}(\mathbf{B}) = k\},$$

and $\lambda_{\min}(\mathbf{B})$ and $\lambda_{\max}(\mathbf{B})$ be the smallest and largest eigenvalues of a square matrix $\mathbf{B}$.

Suppose we have an independent and identically distributed sample $(Y_i^\top,\ X_i^\top)$, $i = 1,\ \cdots,\ n$, from $(Y^\top,\ X^\top)$. A standard penalised least squares estimation would provide us with an estimator $\tilde{\mathbf{A}}$ of $\mathbf{A}$, which is the minimiser of

$$\sum_{i=1}^{n} \|Y_i - \mathbf{A}^\top X_i\|^2 + P_\lambda(\|\mathbf{A}\|_1), \tag{3.1}$$

where $P_\lambda(\cdot)$ is a penalty function. However, this estimation does not take into account the information that $\mathbf{A}$ is of low rank, which would result in an estimator not as efficient as we could expect. In fact, it is easy to see this estimation is equivalent to componentwise penalised least squares estimation for (2.1).

The proposed estimation procedure will fully make use of the low rank information of $\mathbf{A}$, and the resulting estimator will be more efficient.

## 3.1 Estimation method

The idea, based on which the proposed estimation is constructed, is that each column of $\mathbf{A}$ is a linear combination of $r$ linearly independent columns of $\mathbf{A}$. Based on this idea, we propose the following estimation procedure for $\mathbf{A}$. We start with the case when the rank $r$ of $\mathbf{A}$ is known, then propose an estimation for $r$.

### 3.1.1 When $r$ is known

Let $Y_i = (y_{i1},\ \cdots,\ y_{iq})^\top, \quad X_i = (x_{i1},\ \cdots,\ x_{ip})^\top$ for $i = 1,\ \cdots,\ n$. Apply the idea of penalised least squares estimation and minimise

$$\sum_{i=1}^{n} \left\{ \sum_{l=1}^{r} \left( y_{ij_l} - X_i^\top \mathbf{a}_{j_l} \right)^2 + \sum_{k \notin \{j_1,\ \cdots,\ j_r\}} \left( y_{ik} - X_i^\top \sum_{\ell=1}^{r} b_{k\ell} \mathbf{a}_{j_\ell} \right)^2 \right\}$$
$$+ \sum_{l=1}^{r} P_\lambda(\|\mathbf{a}_{j_l}\|_1) \tag{3.2}$$

with respect to $\mathbf{a}_{j_l}$s, $b_{k\ell}$s, and $\{j_1,\ \cdots,\ j_r\}$, where $1 \le j_1 < \cdots < j_r \le q$, $P_\lambda(\cdot)$ is a penalty function, $\lambda$ involved is a tuning parameter which can be selected by some criterion, such as BIC.

When $r$ is large, $b_{k\ell}$s are also likely to have sparsity, in which case, we can add another penalty term into (3.2) to penalise $b_{k\ell}$s. However, when $r$ is small, which is the case of main interest, there is no need to penalise $b_{k\ell}$s, this is because for each component of $Y_i$, say the $k$th component, there are only $r$ $b_{k\ell}$s, which is not many.

Notice that the minimiser of (3.2) is not unique. We denote a minimiser of (3.2) by

$$\hat{\mathcal{D}} = \{\hat{j}_1, \cdots, \hat{j}_r\}, \quad \hat{\mathbf{a}}_j, \ j \in \hat{\mathcal{D}}, \quad \hat{b}_{k\ell}, \ k \in \hat{\mathcal{D}}^c, \ \ell \in \{1, \cdots, r\}.$$

For any $j$, $1 \le j \le q$, the $j$th column of $\mathbf{A}$ is estimated by

$$\begin{cases} \hat{\mathbf{a}}_j, & \text{if } j \in \hat{\mathcal{D}} \\ \sum_{\ell=1}^{r} \hat{b}_{j\ell} \hat{\mathbf{a}}_{\hat{j}_\ell}, & \text{if } j \in \hat{\mathcal{D}}^c \end{cases}$$

We use $\hat{\mathbf{A}}(r)$ to denote the estimator of $\mathbf{A}$.

The non-uniqueness of the minimiser of (3.2) is because that there can be more than one ways to choose $r$ independent columns $\mathcal{D}_r$ so that $\mathbf{A}_{\mathcal{D}_r}$ is full-rank. Theoretically speaking, as long as $\mathbf{A}_{\mathcal{D}_r}$ and $r$ can be well estimated, we can recovery the low-rank structure of $\mathbf{A}$ regardless of the choice of $\mathcal{D}_r$. Our theory shows that the consistency of our proposed estimator holds uniformly in $\mathcal{D}_r$; please refer to Lemmas 1-3 in the Appendix. Hence, this non-unique issue does not affect the performance of the proposed estimation procedure, which is further corroborated via extensive simulations in Section 5.

### 3.1.2 Estimation of $r$

The estimation of $\mathbf{A}$ in section 3.1.1 is built on the assumption that the rank $r$ of $\mathbf{A}$ is known, and this assumption is not realistic in reality. In fact, rank $r$ plays a very important role in the estimation of $\mathbf{A}$. If $r$ is underestimated, a substantial bias would creep into the estimation procedure and make the final estimator of $\mathbf{A}$ very biased. On the other hand, if $r$ is overestimated, we would have to estimate unnecessarily many unknown parameters, which would make

the final estimator of $\mathbf{A}$ have big variance. In this paper, we use BIC, which is defined as follows, to estimate $r$

$$\mathrm{BIC}(k) = \|\mathbf{Y} - \mathbf{X}\hat{\mathbf{A}}(k)\|^2 + \sqrt{kp\log p}h_n, \tag{3.3}$$

where $h_n$ is a positive diverging sequence, which can be set to be $\log n$. The estimator of $r$ is given by

$$\hat{r} = \arg\min_{1\le k\le \bar{r}} \mathrm{BIC}(k),$$

where $\bar{r}$ is a pre-specified bound for $r$. The proposed estimator $\hat{\mathbf{A}}(\hat{r})$ of $\mathbf{A}$ is the $\hat{\mathbf{A}}(r)$, obtained in section 3.1.1, with $r$ being replaced by $\hat{r}$.

We will show in Section 4 the proposed BIC estimator $\hat{r}$ enjoys an excellent asymptotic property, say it tends to identify the true model consistently. If the prediction accuracy is our primary concern, we can consider the multifold cross-validation (CV), which tends to select the model with the optimal prediction performance (Zhang, 1993). The data are splitted randomly into $M$ groups of equal sizes (assuming that $n/M$ is an integer for simplicity), $\mathcal{G}_m, m = 1, \ldots, M$. For each $k$, $1 \le k \le q$, let $\hat{\mathbf{A}}_{-m}(k)$ be the estimator of $\mathbf{A}$, obtained by the method in Section 3.1.1 when the rank of $\mathbf{A}$ is $k$, without using the observations of the $m$th group. The cross-validation sum is defined as

$$\mathrm{CV}(k) = \sum_{m=1}^{M}\sum_{i\in\mathcal{G}_m} \|Y_i - \hat{\mathbf{A}}_{-m}^{\top}(k)X_i\|^2. \tag{3.4}$$

The CV estimator of $r$ is taken to be the minimiser of $\mathrm{CV}(k)$.

## 3.2 Computational algorithm

The minimisation of (3.2) can be difficult. We propose an iterative algorithm to solve this problem. The route of our algorithm is: we first minimise (3.2), for given $\{j_1, \cdots, j_r\}$, with respect to $\mathbf{a}_{j_l}$s, $b_{k\ell}$s, and denote the resulting minimum of (3.2) by $F(j_1, \cdots, j_r)$, then minimise $F(j_1, \cdots, j_r)$, with respect to $j_1, \cdots, j_r$. The details of our algorithm are described as follows.

For any given $\{j_1, \cdots, j_r\}$, we minimise (3.2) with respect to $\mathbf{a}_{j_l}$s and $b_{k\ell}$s by the following iterative approach

(1) We minimise

$$\sum_{i=1}^{n}\sum_{l=1}^{r}\left(y_{ij_l}-X_i^\top \mathbf{a}_{j_l}\right)^2+\sum_{l=1}^{r}P_\lambda(\|\mathbf{a}_{j_l}\|_1)$$

with respect to $\mathbf{a}_{j_l}$s, and denote the minimiser by $\mathbf{a}_{j_l}^{(0)}$s. There are many existing methods to do the minimisation in this step, because this is the minimisation for standard penalised least squares estimation.

(2) Minimise

$$\sum_{i=1}^{n}\sum_{k\notin\{j_1,\ \cdots,\ j_r\}}\left(y_{ik}-X_i^\top\sum_{\ell=1}^{r}b_{k\ell}\mathbf{a}_{j_\ell}^{(0)}\right)^2$$

with respect to $b_{k\ell}$s, and denote the minimiser by $b_{k\ell}^{(0)}$s. Clearly, $b_{k\ell}^{(0)}$ enjoys a closed form, therefore, the minimisation in this step is very easy.

(3) Let $\mathbf{a}_{j_l}^{(0)}$s and $b_{k\ell}^{(0)}$s be the initial values, and minimise (3.2) iteratively. Specifically, let $\mathbf{a}_{j_l}^{(k)}$s and $b_{k\ell}^{(k)}$s be the values of $\mathbf{a}_{j_l}$s and $b_{k\ell}$s in the $k$th iteration. Replace the $b_{k\ell}$s in (3.2) by $b_{k\ell}^{(k)}$s and minimise (3.2) with respect to $\mathbf{a}_{j_l}$s, $\mathbf{a}_{j_l}^{(k+1)}$s are taken to the resulting minimiser.

Replace the $\mathbf{a}_{j_l}$s in (3.2) by $\mathbf{a}_{j_l}^{(k+1)}$s and minimise (3.2) with respect to $b_{k\ell}$s, $b_{k\ell}^{(k+1)}$s are taken to the resulting minimiser.

Continue the iteration until convergence, the limits of $\mathbf{a}_{j_l}^{(k)}$s and $b_{k\ell}^{(k)}$s are the minimser of (3.2), and the minimum of (3.2) is denoted by $F(j_1,\ \cdots,\ j_r)$.

A naive approach to minimise $F(j_1,\ \cdots,\ j_r)$, with respect to $j_1,\ \cdots,\ j_r$, would be to compute $F(j_1,\ \cdots,\ j_r)$ for each possible $\{j_1,\ \cdots,\ j_r\}$, where $1\le j_1<\cdots,<j_r\le q$, and the $\{j_1,\ \cdots,\ j_r\}$ which minimises the obtained $F(j_1,\ \cdots,\ j_r)$s is the minimiser $\{\hat{j}_1,\ \cdots,\ \hat{j}_r\}$ of $F(j_1,\ \cdots,\ j_r)$. However, this approach would have to compute $\binom{q}{r}$ $F(j_1,\ \cdots,\ j_r)$s, which is computationally too expensive. We shall borrow the idea of forward selection to minimise $F(j_1,\ \cdots,\ j_r)$, which is depicted as follows

(I) Let $F(j_1)$ be $F(j_1,\ \cdots,\ j_r)$ when $r=1$, and compute $F(j_1)$ for each possible $j_1$, $1\le j_1\le q$. Let $\hat{j}_1$ be the one which minimises $F(j_1)$.

(II) For any $k < r$, when we have $\{\hat{j}_1, \cdots, \hat{j}_k\}$, the way to select a $j_{k+1}$ from $\{\hat{j}_1, \cdots, \hat{j}_k\}^c$, the set $\{1, \cdots, q\} - \{\hat{j}_1, \cdots, \hat{j}_k\}$, to add into the set $\{\hat{j}_1, \cdots, \hat{j}_k\}$ is as follows: for each possible $j_{k+1}$, we arrange $\hat{j}_1, \cdots, \hat{j}_k$ and $j_{k+1}$ in ascent order, and denote them by $\tilde{j}_1 < \cdots < \tilde{j}_{k+1}$. We compute $F(\tilde{j}_1, \cdots, \tilde{j}_{k+1})$. The selected $j_{k+1}$ is the one which minimises $F(\tilde{j}_1, \cdots, \tilde{j}_{k+1})$. We add the selected $j_{k+1}$ into the set $\{\hat{j}_1, \cdots, \hat{j}_k\}$, and sort the elements in the new set in ascent order. With a little bit abuse of notation, we denote the new set by $\{\hat{j}_1, \cdots, \hat{j}_{k+1}\}$, where $\hat{j}_1 < \cdots < \hat{j}_{k+1}$.

(III) Continue (II) until $k = r$. We use the obtained $\{\hat{j}_1, \cdots, \hat{j}_r\}$ to approximate the minimiser of $F(j_1, \cdots, j_r)$.

Substitute $\{\hat{j}_1, \cdots, \hat{j}_r\}$ for $\{j_1, \cdots, j_r\}$ in (3.2), and minimise (3.2) with respect to $\mathbf{a}_{\hat{j}_l}$s and $b_{k\ell}$s. Denote the resulting minimiser by $\hat{\mathbf{a}}_{\hat{j}_l}$s and $\hat{b}_{k\ell}$s. We take

$$\{\hat{j}_1, \cdots, \hat{j}_r\}, \quad \hat{\mathbf{a}}_{\hat{j}_l}, \ \ l = 1, \cdots, r, \quad \hat{b}_{k\ell}, \ \ k \notin \{\hat{j}_1, \cdots, \hat{j}_r\}, \ \ell = 1, \cdots, r$$

as a minimiser of (3.2) with respect to $\{j_1, \cdots, j_r\}$, $\mathbf{a}_{\hat{j}_l}$s and $b_{k\ell}$s.

# 4 Asymptotic properties

In this section we are going to investigate the asymptotic behavior of the proposed estimator of $\mathbf{A}$.

Throughout this paper, $A_n \sim B_n$ means that there is a constant $C > 1$ such that $B_n/C \leq A_n \leq B_n C$ with probability tending to 1. "$\gtrsim$" and "$\lesssim$" are similarly defined.

To make the theoretical derivation more neat, we write the minimisation of (3.2) in matrix form. Specifically, for any given integer $k \in [1, q)$, when $r = k$, the minimisation of (3.2) can be written to the minimisation of the following objective function

$$\mathcal{L}(\mathbf{U}, \mathbf{V}, \mathcal{D}_k; k) = \|\mathbf{Y}_{\mathcal{D}_k} - \mathbf{X}\mathbf{U}\|^2 + \|\mathbf{Y}_{\mathcal{D}_k^c} - \mathbf{X}\mathbf{U}\mathbf{V}^\top\|^2 + \sum_{\ell=1}^{k} P_{\boldsymbol{\lambda}_n}(\|\mathbf{U}_\ell\|_1) \quad (4.1)$$

with respect to $\{\mathcal{D}_k,\ \mathbf{U} \in \mathbb{G}(k),\ \mathbf{V} \in \mathbb{R}^{(q-k)\times k}\}$.

Without loss of generality, we use the (adaptive) lasso type penalty

$$P_{\boldsymbol{\lambda}_n}(\|\mathbf{U}_\ell\|_1) = n\sum_{j=1}^{p} \lambda_{j\ell}|u_{j\ell}|,$$

see Zou (2006), where $\lambda_{j\ell} > 0$ and $u_{j\ell}$ are the $(j,\ell)$th element of $\mathbf{U}$.

Throughout this section, we assume that each column of $\mathbf{X}$ has been normalized to have $L_2$-norm of $n$. Furthermore, we denote

$$\gamma_{0n} = \min\{\lambda_{j\ell} : a_{j\ell} = 0\}, \quad \gamma_{1n} = \max\{\lambda_{j\ell} : a_{j\ell} \neq 0\},$$

where $a_{j\ell}$ is the $(j,\ell)$th entry of $\mathbf{A}$.

In order to establish the asymptotic properties of the proposed methods, we impose the following technical conditions:

**Condition 1** *There exist positive constants $\bar{\kappa}$ and $\underline{\kappa}$ such that with probability one $\underline{\kappa} \leq \lambda_{\min}(n^{-1}\mathbf{X}^\top\mathbf{X}) < \lambda_{\max}(n^{-1}\mathbf{X}^\top\mathbf{X}) \leq \bar{\kappa}$.*

**Condition 2** *For some positive constants $C$ and $K$, $\mathbb{E}\{\exp(C\epsilon_j^2)\} < K$ for $j = 1, \cdots, q$.*

**Condition 3** *The elements of $\mathbf{A}$ and $\mathbf{XA}$ are bounded. The matrix $\mathbf{A}$ is $s_n$-sparse in the sense that $\max_{1\leq\ell\leq q}\sum_{j=1}^{p}\mathbb{I}(a_{j\ell} \neq 0) \leq s_n$.*

**Condition 4** *$rp\log p/n \to 0$ when $n \to \infty$.*

**Condition 5** *For any given $k \in [1, q)$,*

$$\liminf_{n\to\infty} \frac{\min_{\mathcal{D}\in\bar{\mathbb{H}}_k(r),\mathbf{U}\in\mathbb{G}(k)} \left[\|\mathbf{XU} - \mathbf{XA}_{\mathcal{D}}\|^2 + \mathrm{tr}\left\{(\mathbf{XA}_{\mathcal{D}^c})^\top(\mathbf{I} - \mathbf{H}_{\mathbf{U}})\mathbf{XA}_{\mathcal{D}^c}\right\}\right]}{\max(\sqrt{nrp\log p}, nrs_n\gamma_{1n})} \to \infty,$$

*where $\mathbf{H}_{\mathbf{U}} = \mathbf{Z}_{\mathbf{U}}(\mathbf{Z}_{\mathbf{U}}^\top\mathbf{Z}_{\mathbf{U}})^{-1}\mathbf{Z}_{\mathbf{U}}^\top$ and $\mathbf{Z}_{\mathbf{U}} = n^{-1/2}\mathbf{XU}$.*

**Remark 1** Condition 1 implies that the predictor matrix has a reasonably good behavior; this is a type of restricted eigenvalue assumption and is commonly used in the literature, e.g., Fan and Peng (2004). Condition 2 requires that each entry of $\mathbf{E}$ is sub-Gaussian, which ensures its tail probability decays

exponentially. Condition 3 facilitates our derivation but can be much relaxed so that the true signal strength depends on $n$ as well. Condition 4 imposes requirement on the diverging rate in order to obtain consistent estimation when $p,\ r$ diverges with $n$. Condition 5 is an identifiability assumption, ensuring that a true low-rank structure can be recognized.

We start with the establishment of the asymptotic property of the minimiser $\hat{\mathbf{A}}(r) = (\hat{\mathbf{U}}, \hat{\mathbf{V}})$ of (4.1) when $k = r$. As far as the asymptotic properties are concerned, $p$, $q$ and $r$ are allowed to depend on $n$ and diverge as $n \to \infty$. The reason for us to suppress the subscript $n$ of $p$, $q$ and $r$ is to make notations neat.

We define an index set

$$\mathcal{M}(\mathcal{D}) = \{1 \le j \le p,\ \ell \in \mathcal{D} : a_{j\ell} \neq 0\},$$

and its complement set is denoted by $\mathcal{M}^c(\mathcal{D})$. We have the following theorem:

**Theorem 1** *Under Conditions 1-5, if* $\gamma_{0n}/\sqrt{rp\log p/n} \to \infty$ *and* $rs_n\gamma_{1n}^2 \to 0$ *as* $n \to \infty$*, there exists, with probability tending to one, a local minimiser* $\{\hat{\mathcal{D}}_r, \hat{\mathbf{A}}(r)\}$ *of* $\mathcal{L}(\mathbf{U}, \mathbf{V}, \mathcal{D}_r; r)$ *satisfying:* $\mathbf{A}_{\hat{\mathcal{D}}_r}$ *is full-rank, and*

$$\hat{\mathbf{A}}_{\mathcal{M}^c(\hat{\mathcal{D}}_r)} = 0, \quad \|\hat{\mathbf{A}}(r) - \mathbf{A}\| = O_p(\{r(s_n + q)(\log p/n + \gamma_{1n}^2)\}^{1/2}).$$

Theorem 1 implies that we can identify a "correct" $\hat{\mathcal{D}}_r$ in the sense that $\mathbf{A}_{\hat{\mathcal{D}}_r}$ is full rank with an overwhelming probability when $n$ is large. The penalised estimators of the zero coefficients are exactly zero under some conditions on $\gamma_{0n}$. The condition $rs_n\gamma_{1n}^2 \to 0$ together with Condition 4 ensures that the proposed estimator is consistent. From the proof of this theorem in the Appendix, we can see, as a special case, that $\|\hat{\mathbf{A}}(r) - \mathbf{A}\| = O_p(\sqrt{r(p+q)\log p/n})$ when $\gamma_{0n} = \gamma_{1n} = 0$. This is in line with the relevant existing results, see, for example, Negahban and Wainwright (2011).

When the coefficient matrix is sparse, under properly selected tuning parameters, the proposed penalised estimator would enjoy the "oracle property". Specifically, if the adaptive lasso penalty with tuning parameter $\lambda_{j\ell} = \lambda_n \tilde{a}_{j\ell}^{-1}$ is used, where $\tilde{\mathbf{A}} = (\tilde{a}_{jl})_{p\times q}$ is the standard least-squares estimator of $\mathbf{A}$, we can verify that $\|\hat{\mathbf{A}}(r) - \mathbf{A}\| = O_p(\{r(s_n + q)(\log p/n)\}^{1/2})$, provided that

$rqp^2 \log p/n \to 0$ as $n \to \infty$, by setting $\lambda_n \sim \sqrt{\log p/n}$ and using the fact that $\tilde{a}_{j\ell} = O_p(1)$ for $a_{j\ell} \neq 0$ and $\tilde{a}_{j\ell} = O_p(\sqrt{pq \log p/n})$ for $a_{j\ell} = 0$.

As mentioned before, the estimation of the rank $r$ of $\mathbf{A}$ plays a very important role in the estimation procedure of $\mathbf{A}$. Theorem 2 shows that the proposed BIC estimator $\hat{r}$, defined by (3.3), is consistent.

**Theorem 2** *Under the conditions in Theorem 1, if* $\bar{r}/h_n \to 0$ *as* $n \to \infty$, *we have* $\Pr(\hat{r} = r) \to 1$ *as* $n \to \infty$.

It is very easy to see Theorem 1 together with Theorem 2 imply the proposed estimator of $\mathbf{A}$ is consistent.

## 5 Simulation studies

In this section, we use two simulated examples, low and high dimensional ones, to assess the coefficient matrix estimation, the prediction, and the rank recovery performance of the proposed method. We assess the proposed method and consider the rank being estimated by either 5 fold cross-validation (OUR-CV; Eq.(3.4)) or BIC (OUR-BIC; Eq.(3.3)). Throughout this section, OUR-CV and OUR-BIC are implemented by the algorithm introduced in Section 3.2 with $P_\lambda(\cdot)$ being the $L_1$ penalty function. The tuning parameter $\lambda$ is selected by 5 fold cross-validation.

In the low dimensional example, we compare OUR-CV and OUR-BIC with the Factor Estimation and Selection method (FES) proposed in Yuan et al. (2007), the Rank Selection Criterion (RSC) proposed in Bunea et al. (2011), and the Self-Tuning Rank Selection (STRS) proposed in Bing and Wegkamp (2019). Besides, we also consider the Ordinary Least Squares estimator (OLS) and a Low-rank Matrix Decomposition estimator (LMD) as two benchmarks. Denote $\hat{\mathbf{A}}_{OLS}$ the OLS estimator of $\mathbf{A}$, the LMD estimator of $\mathbf{A}$ is obtained from a rank $\hat{r}$ truncated singular value decomposition of $\hat{\mathbf{A}}_{OLS}$ as

$$\hat{\mathbf{A}}_{LMD} = \mathbf{U}\mathbf{D}\mathbf{V}^\top = \sum_{l=1}^{\hat{r}} \sigma_l \mathbf{u}_l \mathbf{v}_l^\top,$$

where $\mathbf{D} = \text{diag}\{\sigma_1, \ \ldots, \ \sigma_{\hat{r}}\}$ is a diagonal matrix of $\hat{r}$ largest positive singular values of $\hat{\mathbf{A}}_{OLS}$, and $\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_{\widetilde{r}})$ and $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_{\hat{r}})$ are corresponding left and right singular vectors of $\hat{\mathbf{A}}_{OLS}$, respectively. Further, $\hat{r}$ is estimated by the eigen-ratio method, see Ahn and Horenstein (2013) and references therein. In the high dimensional example, we only compare OUR-CV and OUR-BIC with RSC and STRS as FES, OLS and LMD are not applicable.

### 5.1 Simulation settings

Consider the multivariate linear regression model (2.2). Similar to Bunea et al. (2011) and Bing and Wegkamp (2019), we consider a data generating process as follows.

(1) Coefficient matrix $\mathbf{A}$: Let $\mathbf{A} = b\mathbf{\Gamma}_0\mathbf{\Gamma}_1$, with $b > 0$, $\mathbf{\Gamma}_0 \in \mathbb{R}^{p\times r}$, $\mathbf{\Gamma}_1 \in \mathbb{R}^{r\times q}$ and $r \leq \min(p, \ q)$. The entries of $\mathbf{\Gamma}_0$ and $\mathbf{\Gamma}_1$ are independently drawn from $N(0, \ 1)$. The parameter $r$ controls the rank of $\mathbf{A}$. The parameters $b$ and $r$ together control the signal to noise ratio in (2.2).

(2) Error matrix $\mathbf{E}$: The entries of $\mathbf{E}$ are independently drawn from $N(0, \ 1)$.

The design matrix $\mathbf{X}$ is generated with the following two settings to cover the low and high dimensional cases, respectively.

(3a) Design matrix $\mathbf{X}$ when $n > p$ (low dimensional case): $X_i$, $i = 1, \ \cdots, \ n$ are independently drawn from multivariate Normal distribution $N_p(\mathbf{0}, \ \mathbf{\Sigma})$. For $i, j = 1, \ \cdots, \ p$, the $(i, j)$th entry of $\mathbf{\Sigma}$ is defined as $\Sigma_{ij} = \eta^{|i-j|}$ for some $\eta \in (0, \ 1)$.

(3b) Design matrix $\mathbf{X}$ when $p > n > q$ (high dimensional case): Let $\mathbf{X} = \mathbf{\Lambda}_0\mathbf{\Lambda}_1\mathbf{\Sigma}^{1/2}$, with $\mathbf{\Lambda}_0 \in \mathbb{R}^{n\times q}$, and $\mathbf{\Lambda}_1 \in \mathbb{R}^{q\times p}$. The entries of $\mathbf{\Lambda}_0$ and $\mathbf{\Lambda}_1$ are independently drawn from $N(0, \ 1)$. The covariance matrix $\mathbf{\Sigma}$ is defined as in (3a).

For each case, we generate a testing sample $\{\mathbf{Y}^*, \mathbf{X}^*\}$ of size $n^*$ independent of $\{\mathbf{Y}, \mathbf{X}\}$ to assess the prediction performance of each method. To sum up, the parameters that control the data generating process are listed in Table 1.

Table 1: **Summary of parameters in data generating process**

| Parameter | Description |
|---|---|
| $n$ | Training sample size |
| $n^*$ | Testing sample size |
| $p$ | Dimension of covariate variables |
| $q$ | Dimension of response variables |
| $r$ | Rank of coefficient matrix |
| $b$ | Signal strength parameter |
| $\eta$ | Correlation level among covariates |

### 5.2 Low dimensional example

In the first example, we examine the performance of OUR-CV, OUR-BIC, RSC, STRS, FES, OLS and LMD in the low dimensional case. We set $n = 100$, $n^* = 50$, $p = 25$ and $q = 20$. Then, we vary the rank $r = 5,\ 10,\ 15$, the signal strength parameter $b = 0.2,\ 0.4$, and the correlation level $\eta = 0.5,\ 0.9$. We simulate 200 replications for each scenario.

For each replication, we calculate two re-scaled Frobenius norms as

$$\Delta_k = \frac{1}{pq}\|\hat{\mathbf{A}}_k - \mathbf{A}\|^2 \quad \text{and} \quad \Gamma_k = \frac{1}{n^*q}\|\mathbf{Y}^* - \mathbf{X}^*\hat{\mathbf{A}}_k\|^2, \quad k = 1,\ \cdots,\ 200, \tag{5.1}$$

where $\hat{\mathbf{A}}_k$ is the estimate of $\mathbf{A}$ in the $k$th replication, and $\{\mathbf{Y}^*, \mathbf{X}^*\}$ is a testing sample of size $n^*$. Then, we calculate the sample mean and sample standard deviation of $\Delta_k$ and $\Gamma_k$ over 200 replications.

We use $\hat{r}_k$ to denote the estimated rank of $\mathbf{A}$ in the $k$th replication, $k = 1,\ \cdots,\ 200$. The estimation accuracy of the rank $r$ is assessed by the correct rank recovery rate which is defined as

$$R = \frac{1}{200}\sum_{k=1}^{200} I(\hat{r}_k = r), \tag{5.2}$$

where $I(\cdot)$ is the indicator function.

The simulation results of the low dimensional example with $b = 0.2$ and 0.4 are presented in Tables 2 and 3, respectively. In most scenarios, OUR-CV

performs slightly better than OUR-BIC but pays a price on the computational cost. When the rank is small (e.g. $r = 5$), OUR-CV, OUR-BIC and STRS can recovery the correct rank with a rate close to 1 and have small estimation and prediction errors. RSC struggles when $\eta$ is large and resultes in low correct recovery rates. LMD suffers when the signal to noise ratio is small. When the rank is moderate or large (e.g. $r = 10$ or 15), the correct rank recovery rates of RSC, STRS and LMD drop. As a result, the estimation and prediction errors of RSC, STRS and LMD are also inflated. Compared with the other methods, OUR-CV and OUR-BIC are less sensitive to rank, correlation level and the signal to noise ratio. In general, FES performs similar as RSC in terms of estimation and prediction, and OLS performs unsatisfactory as it ignores the low rank structure in $\mathbf{A}$.

### 5.3 High dimensional example

In the high dimensional example, we examine the performance of OUR-CV and OUR-BIC and compare them with RSC and STRS. We set $n = 40$, $n^* = 40$, $p = 100$ and $q = 25$. Then, we vary the rank $r = 10,\ 20$, the signal strength parameter $b = 0.2,\ 0.4$, and the correlation level $\eta = 0.5,\ 0.9$. We simulate 200 replications for each scenario. The estimation and prediction errors are still measured by the sample mean and sample standard deviation of the re-scaled Frobenius norms defined in (5.1). The estimation accuracy of rank $r$ is measured by the correct rank recovery rate defined in (5.2).

The simulation results of the high dimensional example with $b = 0.2$ and 0.4 are presented in Tables 4 and 5, respectively. Similar to the low dimensional case, OUR-CV outperforms OUR-BIC in most scenarios. When both $\eta$ and $r$ are large, OUR-CV and OUR-BIC maintains reasonable correct rank recovery rates while RSC and STRS fail to recovery the correct rank in most replications.

Table 2: **Results for the low dimensional example with $b = 0.2$**

| Rank | Method | $\eta = 0.5$ | | | $\eta = 0.9$ | | |
|---|---|---|---|---|---|---|---|
| | | $\Delta$ | $\Gamma$ | $R$ | $\Delta$ | $\Gamma$ | $R$ |
| 5 | OUR-CV | 0.013(0.005) | 1.121(0.052) | 1.00 | 0.040(0.006) | 1.130(0.053) | 1.00 |
| | OUR-BIC | 0.013(0.005) | 1.124(0.054) | 1.00 | 0.041(0.006) | 1.132(0.053) | 0.99 |
| | RSC | 0.017(0.011) | 1.130(0.074) | 0.87 | 0.062(0.018) | 1.249(0.071) | 0.59 |
| | STRS | 0.013(0.010) | 1.128(0.053) | 1.00 | 0.052(0.013) | 1.133(0.055) | 0.97 |
| | FES | 0.018(0.015) | 1.132(0.045) | NA | 0.073(0.022) | 1.221(0.048) | NA |
| | OLS | 0.022(0.026) | 1.332(0.066) | NA | 0.125(0.027) | 1.335(0.069) | NA |
| | LMD | 0.014(0.008) | 1.253(0.195) | 0.76 | 0.046(0.015) | 1.247(0.075) | 0.68 |
| 10 | OUR-CV | 0.015(0.006) | 1.235(0.058) | 0.98 | 0.087(0.010) | 1.254(0.060) | 0.95 |
| | OUR-BIC | 0.015(0.006) | 1.236(0.058) | 0.96 | 0.089(0.009) | 1.257(0.061) | 0.92 |
| | RSC | 0.018(0.017) | 1.290(0.087) | 0.81 | 0.097(0.019) | 1.278(0.068) | 0.40 |
| | STRS | 0.016(0.013) | 1.244(0.061) | 0.90 | 0.093(0.016) | 1.254(0.059) | 0.65 |
| | FES | 0.019(0.016) | 1.267(0.049) | NA | 0.106(0.022) | 1.276(0.051) | NA |
| | OLS | 0.022(0.028) | 1.341(0.067) | NA | 0.124(0.035) | 1.341(0.067) | NA |
| | LMD | 0.020(0.032) | 1.391(0.230) | 0.38 | 0.135(0.048) | 1.505(0.187) | 0.27 |
| 15 | OUR-CV | 0.020(0.009) | 1.312(0.061) | 0.96 | 0.095(0.012) | 1.321(0.064) | 0.88 |
| | OUR-BIC | 0.020(0.009) | 1.314(0.063) | 0.92 | 0.097(0.012) | 1.322(0.064) | 0.86 |
| | RSC | 0.022(0.019) | 1.338(0.069) | 0.44 | 0.117(0.015) | 1.329(0.067) | 0.28 |
| | STRS | 0.021(0.012) | 1.323(0.067) | 0.65 | 0.104(0.013) | 1.325(0.064) | 0.42 |
| | FES | 0.022(0.015) | 1.326(0.058) | NA | 0.117(0.018) | 1.332(0.059) | NA |
| | OLS | 0.024(0.029) | 1.337(0.065) | NA | 0.124(0.032) | 1.338(0.041) | NA |
| | LMD | 0.029(0.033) | 1.373(0.292) | 0.15 | 0.136(0.041) | 2.012(0.403) | 0.12 |

*The columns $\Delta$ and $\Gamma$ report the sample mean and sample standard deviation (in parentheses) of $\Delta_k$ and $\Gamma_k$, which are defined in* (5.1)*, over 200 replications. The column $R$ reports the rank recovery rate, which is defined in* (5.2)*, over 200 replications.*

# 6 Real data analysis

Particulate matter up to $2.5\mu m$ (PM2.5) is a complex mixture of solid particles, chemicals (e.g. sulfates, nitrates) and liquid droplets in the air, which include inhalable particles that are small enough to penetrate the thoracic region of the respiratory system. Short term (days) exposure to inhalable PM2.5 can cause an increase in hospital admissions related to respiratory and cardiovascular morbidity, such as aggravation of asthma, respiratory symptoms, and cardiovascular disorders. Long term (years) exposure to inhalable PM2.5 may

Table 3: **Results for the low dimensional example with $b = 0.4$**

| Rank | Method | $\eta = 0.5$ | | | $\eta = 0.9$ | | |
|---|---|---|---|---|---|---|---|
| | | $\Delta$ | $\Gamma$ | $R$ | $\Delta$ | $\Gamma$ | $R$ |
| 5 | OUR-CV | 0.010(0.003) | 1.125(0.052) | 1.00 | 0.037(0.005) | 1.125(0.052) | 1.00 |
| | OUR-BIC | 0.011(0.003) | 1.125(0.052) | 0.99 | 0.037(0.005) | 1.126(0.052) | 0.98 |
| | RSC | 0.015(0.005) | 1.168(0.069) | 0.32 | 0.051(0.007) | 1.135(0.056) | 0.82 |
| | STRS | 0.011(0.003) | 1.126(0.052) | 1.00 | 0.037(0.006) | 1.125(0.052) | 1.00 |
| | FES | 0.015(0.008) | 1.154(0.026) | NA | 0.054(0.013) | 1.139(0.051) | NA |
| | OLS | 0.023(0.015) | 1.332(0.066) | NA | 0.125(0.027) | 1.332(0.066) | NA |
| | LMD | 0.011(0.004) | 1.171(0.095) | 0.99 | 0.041(0.017) | 1.280(0.085) | 0.86 |
| 10 | OUR-CV | 0.014(0.006) | 1.231(0.057) | 1.00 | 0.078(0.009) | 1.238(0.059) | 0.97 |
| | OUR-BIC | 0.014(0.006) | 1.231(0.057) | 0.99 | 0.080(0.009) | 1.240(0.059) | 0.95 |
| | RSC | 0.019(0.017) | 1.309(0.069) | 0.86 | 0.093(0.017) | 1.292(0.078) | 0.66 |
| | STRS | 0.015(0.013) | 1.233(0.057) | 0.98 | 0.082(0.011) | 1.245(0.061) | 0.73 |
| | FES | 0.017(0.016) | 1.312(0.055) | NA | 0.096(0.012) | 1.258(0.060) | NA |
| | OLS | 0.024(0.021) | 1.340(0.027) | NA | 0.124(0.025) | 1.341(0.067) | NA |
| | LMD | 0.018(0.016) | 1.322(0.067) | 0.90 | 0.108(0.035) | 1.506(0.109) | 0.45 |
| 15 | OUR-CV | 0.019(0.009) | 1.301(0.061) | 0.97 | 0.103(0.011) | 1.314(0.065) | 0.91 |
| | OUR-BIC | 0.020(0.009) | 1.303(0.062) | 0.95 | 0.104(0.011) | 1.315(0.065) | 0.88 |
| | RSC | 0.021(0.019) | 1.331(0.071) | 0.63 | 0.119(0.014) | 1.339(0.071) | 0.35 |
| | STRS | 0.021(0.012) | 1.322(0.069) | 0.81 | 0.107(0.011) | 1.323(0.067) | 0.54 |
| | FES | 0.021(0.015) | 1.327(0.063) | NA | 0.110(0.013) | 1.327(0.066) | NA |
| | OLS | 0.022(0.023) | 1.335(0.065) | NA | 0.125(0.025) | 1.342(0.065) | NA |
| | LMD | 0.028(0.037) | 1.414(0.165) | 0.36 | 0.175(0.097) | 1.949(0.269) | 0.17 |

*The columns $\Delta$ and $\Gamma$ report the sample mean and sample standard deviation (in parentheses) of $\Delta_k$ and $\Gamma_k$, which are defined in* (5.1)*, over 200 replications. The column $R$ reports the rank recovery rate, which is defined in* (5.2)*, over 200 replications.*

lead to an increase in mortality from cardiovascular and respiratory diseases, like lung cancer. The hazardous effects of inhalable PM2.5 on human health have been well-documented, see Riediker et al. (2004), Polichetti et al. (2009), Franck et al. (2011), Xing et al. (2016), Pun et al. (2017) and references therein.

In this section, we investigate the relationship between concentration of PM2.5 and four air pollutants: ozone, sulfur dioxide (SO2), carbon monoxide (CO), and nitrogen dioxide (NO2). The dataset for us to study is available at

`https://www.epa.gov/outdoor-air-quality-data`,

Table 4: **Results for the high dimensional example with $b = 0.2$**

| | | $\eta = 0.5$ | | | $\eta = 0.9$ | | |
|---|---|---|---|---|---|---|---|
| Rank | Method | $\Delta$ | $\Gamma$ | $R$ | $\Delta$ | $\Gamma$ | $R$ |
| 10 | OUR-CV | 0.225(0.030) | 1.468(0.142) | 1.00 | 0.273(0.034) | 1.508(0.146) | 0.96 |
| | OUR-BIC | 0.226(0.031) | 1.471(0.146) | 0.99 | 0.273(0.034) | 1.511(0.146) | 0.96 |
| | RSC | 0.299(0.036) | 2.165(0.185) | 0.85 | 0.351(0.085) | 2.392(0.203) | 0.38 |
| | STRS | 0.226(0.031) | 1.470(0.144) | 0.97 | 0.2857(0.039) | 1.522(0.151) | 0.94 |
| 20 | OUR-CV | 0.402(0.047) | 1.679(0.183) | 0.93 | 0.502(0.046) | 1.790(0.208) | 0.85 |
| | OUR-BIC | 0.406(0.049) | 1.685(0.184) | 0.91 | 0.508(0.048) | 1.794(0.210) | 0.84 |
| | RSC | 0.618(0.072) | 2.602(0.471) | 0.15 | 0.758(0.091) | 2.627(0.616) | 0.06 |
| | STRS | 0.449(0.058) | 2.052(0.235) | 0.62 | 0.561(0.062) | 2.156(0.273) | 0.51 |

*The columns $\Delta$ and $\Gamma$ report the sample mean and sample standard deviation (in parentheses) of $\Delta_k$ and $\Gamma_k$, which are defined in* (5.1)*, over 200 replications. The column $R$ reports the rank recovery rate, which is defined in* (5.2)*, over 200 replications.*

Table 5: **Results for the high dimensional example with $b = 0.4$**

| | | $\eta = 0.5$ | | | $\eta = 0.9$ | | |
|---|---|---|---|---|---|---|---|
| Rank | Method | $\Delta$ | $\Gamma$ | $R$ | $\Delta$ | $\Gamma$ | $R$ |
| 10 | OUR-CV | 0.949(0.102) | 1.526(0.148) | 1.00 | 0.977(0.122) | 1.560(0.151) | 0.99 |
| | OUR-BIC | 0.951(0.104) | 1.527(0.148) | 0.98 | 0.979(0.123) | 1.568(0.153) | 0.98 |
| | RSC | 1.198(0.124) | 2.392(0.188) | 0.74 | 1.201(0.159) | 2.548(0.195) | 0.50 |
| | STRS | 0.954(0.107) | 1.528(0.150) | 1.00 | 0.982(0.126) | 1.570(0.156) | 0.98 |
| 20 | OUR-CV | 1.021(0.152) | 1.770(0.201) | 0.95 | 1.181(0.184) | 1.838(0.211) | 0.88 |
| | OUR-BIC | 1.025(0.155) | 1.774(0.202) | 0.92 | 1.183(0.188) | 1.841(0.214) | 0.87 |
| | RSC | 1.808(0.406) | 2.656(0.482) | 0.35 | 2.059(0.685) | 2.744(0.658) | 0.13 |
| | STRS | 1.457(0.189) | 2.104(0.262) | 0.78 | 1.513(0.197) | 2.210(0.285) | 0.67 |

*The columns $\Delta$ and $\Gamma$ report the sample mean and sample standard deviation (in parentheses) of $\Delta_k$ and $\Gamma_k$, which are defined in* (5.1)*, over 200 replications. The column $R$ reports the rank recovery rate, which is defined in* (5.2)*, over 200 replications.*

it was collected from 37 outdoor monitoring sites across the United States. Specifically, the concentration of each of the 4 pollutants was measured and collected daily from the 37 sites between January 2017 to April 2019, and it has 729 observations in total. The concentration of PM2.5 was collected in the

same manner.

What we are interested in is the association between the concentrations of PM2.5 and the concentrations of the four air pollutants at the 37 monitor sites. As the concentrations of the four air pollutants at one site may also contribute the concentrations of PM2.5 at other sites, we include the concentrations of the four air pollutants at all 37 sites in the explanatory variables for the concentration of PM2.5 at each site of the 37 sites, this gives us 148 explanatory variables for the concentration of PM2.5 at each site. In Figure 1, we plot the sample means of the concentrations of PM2.5 and of the four air pollutants against the geological locations where they were collected.

We take the first-order difference for each column of the dataset to remove the non-stationarity, and standardize it to make it have mean 0 and variance 1. Let $\mathbf{Y} = (\mathbf{Y}^1, \cdots, \mathbf{Y}^{37}) \in \mathbb{R}^{728\times 37}$ be the matrix of the 728 observations of the response variable which contains the concentrations of PM2.5 collected by the 37 sites, and $\mathbf{X} = (\mathbf{X}^1, \cdots, \mathbf{X}^{148}) \in \mathbb{R}^{728\times 148}$ be matrix of the 728 observations of predictor which contains the 148 explanatory variables. We apply the multivariate linear regression model (2.2) to fit the dataset, where $\mathbf{E} \in \mathbb{R}^{728\times 37}$ is the matrix of random errors, and $\mathbf{A} \in \mathbb{R}^{148\times 37}$ is the coefficient matrix of interest.

We compare the prediction performance of our method with rank estimated by the 5-fold cross-validation (OUR-CV), the Self-Tuning Rank Selection (STRS) proposed in Bing and Wegkamp (2019), and the Ordinary Least Squares estimator (OLS). For each method, we use the first 600 observations as the training set and predict the remaining 128 observations (the test set). Let $\mathbf{Y}_{test}$ and $\hat{\mathbf{Y}}_{test}$ be the matrices respectively of true and predicted values (obtained by one of the three methods listed above) of the response variable in the test set. The prediction accuracy is measured by the mean squared Frobenius norm of the difference between $\hat{\mathbf{Y}}_{test}$ and $\mathbf{Y}_{test}$, which is defined as

$$\text{Prediction Error} = \frac{1}{n_t q}\|\hat{\mathbf{Y}}_{test} - \mathbf{Y}_{test}\|^2,$$

where $q = 37$ and $n_t = 128$, which are the number of columns and the number of rows of $\mathbf{Y}_{test}$, respectively.

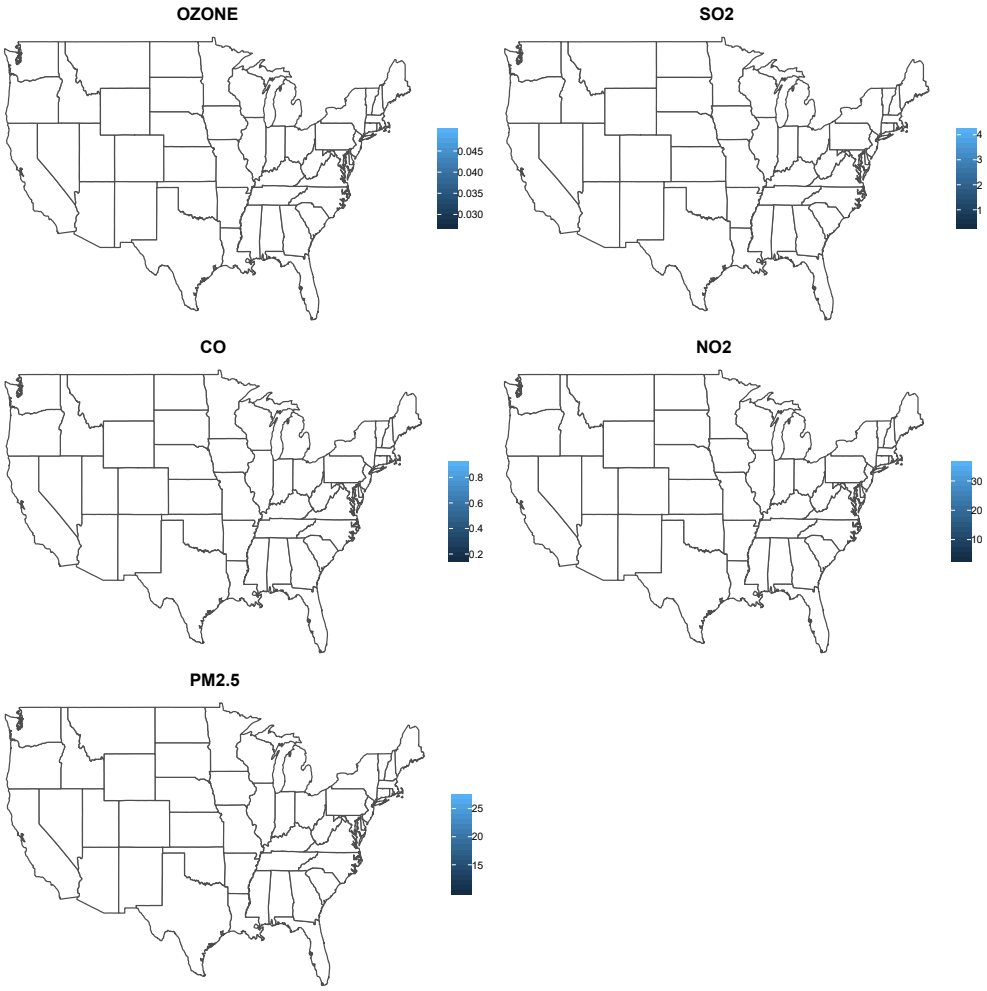


Figure 1: Sample means of the concentrations of PM2.5 and of the four air pollutants

We report in Table 6 the prediction error as well as the estimated rank of the coefficient matrix for each method concerned. According to Table 6, OUR-CV achieves the smallest prediction error among the three competitors. Besides, the rank estimated by OUR-CV is 3 which is more parsimonious than the one estimated by STRS. To justify the rank estimation results, we apply eigen-decomposition to the coefficient matrix estimated by the OLS method, and draw the scree plot with the top 20 eigenvalues in Figure 2. The scree plot shows a clear elbow shape at the third eigenvalue.

Table 6: **Prediction error and estimated rank of the coefficient matrix.**

| Methods | OUR-CV | STRS | OLS |
|---|---|---|---|
| Prediction error | 0.7692 | 0.8551 | 1.0394 |
| Estimated rank | 3 | 10 | NA |

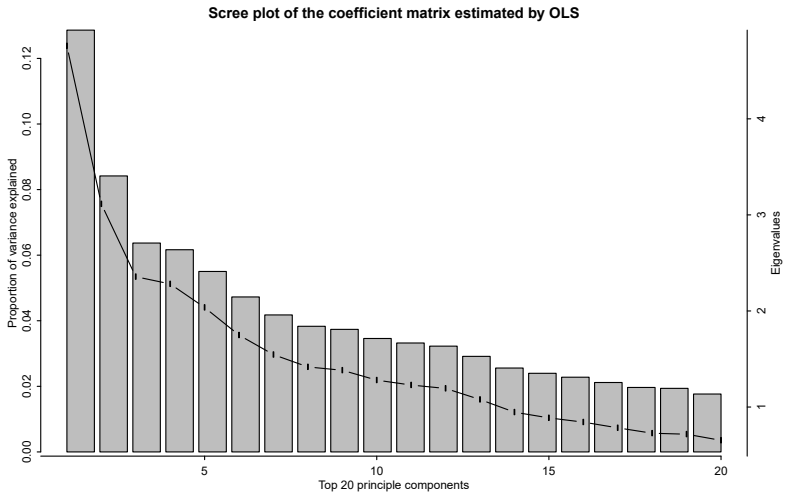


Figure 2: Scree plot of the coefficient matrix estimated by the OLS method. The solid dotted line denotes the leading eigenvalues in descending order. The grey bars denote the proporation of variance explained by each eigenvector.

The coefficient matrix estimated by OUR-CV unveils a parsimonious yet interpretable relationship between PM2.5 and the other four air pollutants. Among the four pollutants, CO has the largest positive contribution to the concentration of PM2.5. As we know, CO is usually produced in the incomplete combustion of carbon-containing fuels, such as gasoline, natural gas, coal, and wood. Two major anthropogenic sources of CO in the United States are vehicle emissions and heating. We find that monitors located in California and around New York City have high CO coefficients which are caused by the dense vehicle population. Also, we notice that the monitors with higher latitudes have higher CO coefficients which may reflect the impact of heating. Further, the attributes

of each pollutant tend to cluster into three geological areas in the United States: west coast, central and east coast.

# Appendix

## Appendix: proofs

Given $\mathcal{D}_k$, the objective function (with respect to $(\mathbf{U}, \mathbf{V})$) is

$$\begin{aligned}\mathcal{L}(\mathbf{U}, \mathbf{V}; \mathcal{D}_k) =& \|\mathbf{Y}_{\mathcal{D}_k} - \mathbf{X}\mathbf{U}\|^2 + \|\mathbf{Y}_{\mathcal{D}_k^c} - \mathbf{X}\mathbf{U}\mathbf{V}^\top\|^2 + n\sum_{\ell=1}^{k}\sum_{j=1}^{p}\lambda_{j\ell}|u_{j\ell}| \\ :=& \mathcal{L}_1(\mathbf{U}; \mathcal{D}_k) + \mathcal{L}_2(\mathbf{U}, \mathbf{V}; \mathcal{D}_k) + n\sum_{\ell=1}^{k}\sum_{j=1}^{p}\lambda_{j\ell}|u_{j\ell}|.\end{aligned}$$

We present three useful lemmas.

**Lemma 1** *Suppose Conditions 1-4 are satisfied. The following result holds uniformly for $\mathcal{D}_r$ such that $\mathbf{A}_{\mathcal{D}_r}$ is full-rank: With probability tending to one, there exists a local minimiser $\hat{\mathbf{A}}(r)$ of $\mathcal{L}(\mathbf{U}, \mathbf{V}; \mathcal{D}_r)$ such that $\|\hat{\mathbf{A}}(r) - \mathbf{A}\| = O_p(\alpha_n)$, where $\alpha_n = \sqrt{r(p+q)\log p/n} + \sqrt{rs_n}\gamma_{1n}$.*

*Proof.* For notational simplicity, in what follows we suppress the dependence of $\mathcal{D}$ on $r$ and write $\mathcal{L}(\mathbf{U}, \mathbf{V}; \mathcal{D}_r)$ as $\mathcal{L}(\mathbf{U})$. It is easy to verify that for a given $\mathbf{U}$, by minimising the function $\mathcal{L}_2(\mathbf{U}, \mathbf{V}; \mathcal{D}_r)$ with respect to $\mathbf{V}$, we obtain that

$$\mathcal{L}_2(\mathbf{U}) := \mathcal{L}_2(\mathbf{U}, \hat{\mathbf{V}}; \mathcal{D}_r) = \operatorname{tr}\left\{\mathbf{Y}_{\mathcal{D}^c}^\top(\mathbf{I} - \mathbf{H}_{\mathbf{U}})\mathbf{Y}_{\mathcal{D}^c}\right\}. \tag{A.1}$$

We will show that there exists a large constant $C > 0$ such that

$$\Pr\left\{\inf_{\mathbf{W}\in\mathbb{R}^{p\times r};\|\mathbf{W}\|=C}\mathcal{L}(\mathbf{A}_{\mathcal{D}} + \alpha_n\mathbf{W}) < \mathcal{L}(\mathbf{A}_{\mathcal{D}}),\ \forall\mathcal{D} \in \mathbb{H}_r(r)\right\} \to 0, \tag{A.2}$$

which implies with probability tending to one that there exists a local minimum in the ball $\{\mathbf{A}_{\mathcal{D}} + \alpha_n\mathbf{W} : \|\mathbf{W}\| \le C\}$ uniformly in $\mathcal{D}$. Hence, there exists a local minimiser of $\mathcal{L}(\mathbf{U})$ such that $\|\hat{\mathbf{U}} - \mathbf{A}_{\mathcal{D}}\| = O_p(\alpha_n)$.

Write

$$\begin{aligned}
&\mathcal{L}(\mathbf{A}_{\mathcal{D}} + \alpha_n \mathbf{w}) - \mathcal{L}(\mathbf{A}_{\mathcal{D}}) \\
=&\{\mathcal{L}_1(\mathbf{A}_{\mathcal{D}} + \alpha_n \mathbf{w}) - \mathcal{L}_1(\mathbf{A}_{\mathcal{D}})\} + \{\mathcal{L}_2(\mathbf{A}_{\mathcal{D}} + \alpha_n \mathbf{w}) - \mathcal{L}_2(\mathbf{A}_{\mathcal{D}})\} \\
&+ n \sum_{\ell \in \mathcal{D}} \sum_{j=1}^{p} \lambda_{j\ell}(|a_{j\ell} + \alpha_n w_{j\ell}| - |a_{j\ell}|) \\
:=&\Delta_1 + \Delta_2 + \Delta_3.
\end{aligned}$$

Observe that $\Delta_1 = n\alpha_n^2 \mathrm{tr}\{\mathbf{w}^\top (n^{-1}\mathbf{X}^\top \mathbf{X})\mathbf{w}\} - 2\alpha_n \mathrm{tr}(\mathbf{E}_{\mathcal{D}}^\top \mathbf{X}\mathbf{w})$. By Condition 1, we have $\mathrm{tr}\{\mathbf{w}^\top (n^{-1}\mathbf{X}^\top \mathbf{X})\mathbf{w}\} \geq \underline{\kappa}\|\mathbf{w}\|^2$. It follows then, that the first term of $\Delta_1$ is uniformly larger than $C^2 \underline{\kappa} n \alpha_n^2$, which is quadratic in $C$.

For the second term of $\Delta_1$, using Cauchy-Schwartz inequality, we have

$$2\alpha_n \mathrm{tr}(\mathbf{E}_{\mathcal{D}}^\top \mathbf{X}\mathbf{w}) \leq 2\alpha_n C \left( \sum_{\ell \in \mathcal{D}} \sum_{j=1}^{p} (\mathbf{X}_j^\top \mathbf{E}^\ell)^2 \right)^{1/2}.$$

By Condition 2 and tail probability of sub-Gaussian variables, we have

$$pq \Pr\left( |\mathbf{X}_j^\top \mathbf{E}^\ell| > c\sqrt{n \log p} \right) \to 0 \tag{A.3}$$

for a sufficiently large $c > 0$, and thus $\mathbf{X}_j^\top \mathbf{E}^\ell = O_p(\sqrt{n \log p})$ uniformly in $j$ and $\ell$. Consequently, the second term is uniformly of order $O_p(\sqrt{rnp \log p}\alpha_n C)$, which is linear in $C$. Therefore, by the definition of $\alpha_n$, as long as the constant $C$ is sufficiently large, the first term dominates the second term with arbitrarily large probability.

Next, we deal with $\Delta_2$. To facilitate the presentation, denote $\tilde{\mathbf{A}}_{\mathcal{D}} = \mathbf{A}_{\mathcal{D}} + \alpha_n \mathbf{w}$, $\tilde{\mathbf{Z}} = n^{-1/2}\mathbf{X}\tilde{\mathbf{A}}_{\mathcal{D}}$, $\mathbf{H}_{\mathbf{A}_{\mathcal{D}}} = \mathbf{H}$ and $\tilde{\mathbf{H}}_{\tilde{\mathbf{A}}_{\mathcal{D}}} = \tilde{\mathbf{H}}$.

By $\mathbf{Y}_{\mathcal{D}^c} = \mathbf{X}\mathbf{A}_{\mathcal{D}}\mathbf{V}^{*\top} + \mathbf{E}_{\mathcal{D}^c}$ and $\mathbf{Z}(\mathbf{I} - \mathbf{H}) = \mathbf{0}$, simple algebra yields that

$$\begin{aligned}
\Delta_2 = &\alpha_n^2 \mathrm{tr}\left\{ (\mathbf{X}\mathbf{w}\mathbf{V}^{*\top})^\top (\mathbf{I} - \tilde{\mathbf{H}})(\mathbf{X}\mathbf{w}\mathbf{V}^{*\top}) \right\} + 2\alpha_n \mathrm{tr}\left\{ \mathbf{E}_{\mathcal{D}^c}^\top (\mathbf{I} - \tilde{\mathbf{H}})(\mathbf{X}\mathbf{w}\mathbf{V}^{*\top}) \right\} \\
&+ \mathrm{tr}\left\{ \mathbf{E}_{\mathcal{D}^c}^\top (\mathbf{H} - \tilde{\mathbf{H}})\mathbf{E}_{\mathcal{D}^c} \right\},
\end{aligned} \tag{A.4}$$

where $\mathbf{V}^* \in \mathbb{R}^{(q-r)\times r}$ such that $\mathbf{A}_{\mathcal{D}}\mathbf{V}^{*\top} = \mathbf{A}_{\mathcal{D}^c}$.

Following the same arguments as in $\Delta_1$, it can be shown that as long as the constant $C$ is sufficiently large, the first term on the right side of (A.4) will

always dominate the second term with arbitrarily large probability. Consider the third term on the right side of (A.4). By Condition 1 again, we get

$$\begin{aligned}\|\tilde{\mathbf{Z}} - \mathbf{Z}\| \leq \bar{\kappa}\|\tilde{\mathbf{A}}_{\mathcal{D}} - \mathbf{A}_{\mathcal{D}}\| &= O(\alpha_n C),\\ \|\tilde{\mathbf{Z}}^\top\tilde{\mathbf{Z}} - \mathbf{Z}^\top\mathbf{Z}\| &= O(\alpha_n C),\\ \|(\tilde{\mathbf{Z}}^\top\tilde{\mathbf{Z}})^{-1} - (\mathbf{Z}^\top\mathbf{Z})^{-1}\| &= O(\alpha_n C),\end{aligned}$$

and accordingly $\|\tilde{\mathbf{H}} - \mathbf{H}\| = O(\alpha_n C)$. By (A.3), we have that

$$\begin{aligned}&\operatorname{tr}\left\{\mathbf{E}_{\mathcal{D}^c}^\top(\mathbf{H} - \tilde{\mathbf{H}})\mathbf{E}_{\mathcal{D}^c}\right\}\\ &= \operatorname{tr}\left[n^{-1}\mathbf{E}_{\mathcal{D}^c}^\top\mathbf{X}\left\{\mathbf{A}_{\mathcal{D}}(\mathbf{Z}^\top\mathbf{Z})^{-1}(\mathbf{A}_{\mathcal{D}})^\top - \tilde{\mathbf{A}}_{\mathcal{D}}(\tilde{\mathbf{Z}}^\top\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{A}}_{\mathcal{D}}^\top\right\}\mathbf{X}^\top\mathbf{E}_{\mathcal{D}^c}\right]\\ &= \operatorname{tr}\left[n^{-1/2}\mathbf{E}_{\mathcal{D}^c}^\top\mathbf{X}\left\{\mathbf{A}_{\mathcal{D}}(\mathbf{Z}^\top\mathbf{Z})^{-1}(\mathbf{A}_{\mathcal{D}})^\top - \tilde{\mathbf{A}}_{\mathcal{D}}(\tilde{\mathbf{Z}}^\top\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{A}}_{\mathcal{D}}^\top\right\}n^{-1/2}\mathbf{X}^\top\mathbf{E}_{\mathcal{D}^c}\right]\\ &\leq O(\alpha_n C)\|n^{-1/2}\mathbf{E}_{\mathcal{D}^c}^\top\mathbf{X}\|^2\\ &= O_p(\alpha_n C r p\log p)\end{aligned}$$

holds uniformly in $\mathcal{D}$, which is of smaller order of $\Delta_1$.

For $\Delta_3$, observe that

$$\begin{aligned}\Delta_3 &:= \Delta_{31} + \Delta_{32}\\ &= n\sum_{j,\ell\in\mathcal{M}(\mathcal{D})}\sum\lambda_{j\ell}\left(|a_{j\ell} + \alpha_n w_{j\ell}| - |a_{j\ell}|\right) + \Delta_{32},\end{aligned}$$

where $\Delta_{32} \geq 0$. By the definition of $\alpha_n$ and $s_n$, $|\Delta_{31}| \leq \sqrt{rs_n}n\alpha_n\gamma_{1n}\|\mathbf{w}\|$ which is dominated by the $\Delta_1$. Hence, by choosing a sufficiently large $C$, (A.2) holds.

By similar arguments, we can verify that $\|\hat{\mathbf{V}} - \mathbf{V}^*\| = O_p(\alpha_n)$, and accordingly $\|\hat{\mathbf{A}}(r) - \mathbf{A}\| = O_p(\alpha_n)$ follows. □

The next lemma establishes the sparsity property of $\hat{\mathbf{A}}(r)$.

**Lemma 2** *Suppose the conditions given in Lemma 1 all hold. If* $n^{1/2}\gamma_{0n}/\sqrt{rp\log p} \to \infty$, *then the following result holds uniformly for* $\mathcal{D}_r$ *such that* $\mathbf{A}_{\mathcal{D}_r}$ *is full-rank: For any* $\hat{\mathbf{A}}_{\mathcal{D}_r}$ *satisfying* $\|\hat{\mathbf{A}}_{\mathcal{D}_r} - \mathbf{A}_{\mathcal{D}_r}\| = O_p(\alpha_n)$, $\Pr(\hat{a}_{j\ell} = 0, \forall j, \ell \in \mathcal{M}^c(\mathcal{D}_r)) \to 1$, *where* $\hat{\mathbf{A}} = (\hat{a}_{j\ell})_{p\times q}$.

*Proof.* The objective function can be written as

$$\mathcal{L}(\mathbf{U}) = \mathrm{tr}(\mathbf{Y}_{\mathcal{D}}^\top \mathbf{Y}_{\mathcal{D}} + \mathbf{Y}_{\mathcal{D}^c}^\top \mathbf{Y}_{\mathcal{D}^c}) - 2\mathrm{tr}(\mathbf{Y}_{\mathcal{D}}^\top \mathbf{X}\mathbf{U} + \mathbf{Y}_{\mathcal{D}^c}^\top \mathbf{X}\mathbf{U}\mathbf{V}^\top) + \mathrm{tr}(\mathbf{U}^\top \mathbf{X}^\top \mathbf{X}\mathbf{U} + \mathbf{U}^\top \mathbf{X}^\top \mathbf{X}\mathbf{U}\mathbf{V}^\top \mathbf{V}) + n\sum_{\ell=1}^{r}\sum_{j=1}^{p} \lambda_{j\ell}|u_{j\ell}|.$$

The $\hat{\mathbf{U}}^\ell$, the $\ell$th column of $\hat{\mathbf{U}}$, satisfies the KKT optimality condition

$$\frac{\partial \mathcal{L}(\mathbf{U})}{\partial u_{j\ell}}_{|\hat{\mathbf{A}}_{\mathcal{D}}} = 2(\mathbf{X}^\top)_j(\mathbf{X}\hat{\mathbf{A}}_{\mathcal{D}} - \mathbf{Y}_{\mathcal{D}})^\ell + 2(\mathbf{X}^\top)_j(\mathbf{X}\hat{\mathbf{A}}_{\mathcal{D}}\mathbf{V}^\top - \mathbf{Y}_{\mathcal{D}^c})\mathbf{V}^\ell + n\lambda_{j\ell}\mathrm{sgn}(\hat{a}_{j\ell}) = 0. \tag{A.5}$$

Firstly, consider the first term of (A.5). Note that

$$n^{-1/2}(\mathbf{X}^\top)_j(\mathbf{X}\mathbf{A}_{\mathcal{D}} - \mathbf{Y}_{\mathcal{D}})^\ell = -n^{-1/2}(\mathbf{X}^\top)_j\mathbf{E}_{\mathcal{D}}^\ell = O_p(\sqrt{\log p})$$

holds uniformly in $j$ and $\mathcal{D}$. By $\|\hat{\mathbf{A}}_{\mathcal{D}} - \mathbf{A}_{\mathcal{D}}\| = O_p(\alpha_n)$, we have

$$\begin{aligned} &n^{-1/2}(\mathbf{X}^\top)_j(\mathbf{X}\hat{\mathbf{A}}_{\mathcal{D}} - \mathbf{Y}_{\mathcal{D}})^\ell \\ &= n^{-1/2}(\mathbf{X}^\top)_j(\mathbf{X}\mathbf{A}_{\mathcal{D}} - \mathbf{Y}_{\mathcal{D}})^\ell + n^{-1/2}(\mathbf{X}^\top)_j\mathbf{X}(\hat{\mathbf{A}}_{\mathcal{D}} - \mathbf{A}_{\mathcal{D}})^\ell \\ &= O_p(\sqrt{\log p}) + n^{-1/2}(\mathbf{X}^\top)_j\mathbf{X}O_p(\alpha_n) \\ &= O_p(\sqrt{n}\alpha_n). \end{aligned} \tag{A.6}$$

Next, for the second term of (A.5), observe that

$$\begin{aligned} &n^{-1/2}(\mathbf{X}^\top)_j(\mathbf{X}\hat{\mathbf{A}}_{\mathcal{D}}\mathbf{V}^\top - \mathbf{Y}_{\mathcal{D}^c}) \\ &= -n^{-1/2}(\mathbf{X}^\top)_j\mathbf{E}_{\mathcal{D}^c} + n^{-1/2}(\mathbf{X}^\top)_j\mathbf{X}(\hat{\mathbf{A}}_{\mathcal{D}}\mathbf{V}^\top - \mathbf{A}_{\mathcal{D}}\mathbf{V}^{*\top}) \end{aligned}$$

and consequently,

$$n^{-1/2}(\mathbf{X}^\top)_j(\mathbf{X}\hat{\mathbf{A}}_{\mathcal{D}}\mathbf{V}^\top - \mathbf{Y}_{\mathcal{D}^c})\mathbf{V}^\ell = O_p(\sqrt{n}\alpha_n) \tag{A.7}$$

holds uniformly in $j$ and $\mathcal{D}$.

Finally, notice that if $\hat{a}_{j\ell} \neq 0$ for $j, \ell \in \mathcal{M}^c(\mathcal{D}_r)$, then $\mathrm{sgn}(\hat{a}_{j\ell}) \neq 0$. Combining (A.6) and (A.7), and the assumption that $n^{1/2}\gamma_{0n}/\sqrt{rp\log p} \to \infty$, then (A.5) will not hold for any $j, \ell \in \mathcal{M}^c(\mathcal{D}_r)$. This is a contradiction, which yields the assertion of this lemma. □

**Lemma 3** *Suppose Conditions 1-4 are satisfied. The following result holds uniformly for $\mathcal{D}_r$ such that $\mathbf{A}_{\mathcal{D}_r}$ is full-rank: With probability tending to one, there exists a local minimiser $\hat{\mathbf{A}}$ of $\mathcal{L}(\mathbf{U},\mathbf{V};\mathcal{D}_r)$ such that $\hat{\mathbf{A}}_{\mathcal{M}^c(\mathcal{D}_r)}=0$, and $\|\hat{\mathbf{A}}-\mathbf{A}\|=O_p(\beta_{n,k})$, where $\beta_{n,k}=\sqrt{r(s_n+q)(\log p/n+\gamma_{1n}^2)}$.*

*Proof.* By Lemma 2, with probability tending to one, $\hat{\mathbf{A}}_{\mathcal{M}^c(\mathcal{D}_r)}=0$. Hence, it suffices to show that there exists a large constant $C>0$ such that

$$\Pr\left\{\inf_{\|\mathbf{w}\|=C}\mathcal{L}(\mathbf{A}_{\mathcal{D}}+\beta_{n,k}\mathbf{w})<\mathcal{L}(\mathbf{A}_{\mathcal{D}}),\ \forall\mathcal{D}\in\mathbb{H}_r(r)\right\}\to 0, \tag{A.8}$$

where $\mathbf{w}_{\mathcal{M}^c(\mathcal{D})}=0$. The proof of (A.8) follows similarly from the arguments in the proof of Lemma 1, except for the second term of $\Delta_1$. Notice that

$$2\beta_{n,k}\mathrm{tr}(\mathbf{E}_{\mathcal{D}}^\top\mathbf{X}\mathbf{w})=O_p(C\beta_{n,k}\sqrt{rns_n\log p}),$$

where we use (A.3) again. Therefore, all the other arguments in the proof of Lemma 1 follows with $\alpha_n$ replaced with $\beta_n$. □

**Lemma 4** *Suppose Conditions 1-5 are satisfied. With probability tending to one, $\mathbf{A}_{\hat{\mathcal{D}}_r}$ is full-rank, where $(\hat{\mathcal{D}}_r,\hat{\mathbf{U}},\hat{\mathbf{V}})$ is the minimiser of $\mathcal{L}(\mathbf{U},\mathbf{V},\mathcal{D}_r;r)$.*

*Proof.* It suffices to show that there exists some $\mathcal{K}_r\in\mathbb{H}_r$,

$$\Pr\left\{\min_{\mathcal{D}_r\in\bar{\mathbb{H}}_r}\min_{\mathbf{U}}\mathcal{L}(\mathbf{U};\mathcal{D}_r)<\mathcal{L}(\mathbf{A}_{\mathcal{K}_r};\mathcal{K}_r)\right\}\to 0. \tag{A.9}$$

Consider $\mathcal{D}\in\bar{\mathbb{H}}_r$. Firstly, by the proof of Lemma 1, we see that

$$\begin{aligned}\mathcal{L}(\mathbf{U})=&\,\mathrm{tr}(\mathbf{Y}_{\mathcal{D}}^\top\mathbf{Y}_{\mathcal{D}})-2\mathrm{tr}(\mathbf{Y}_{\mathcal{D}}^\top\mathbf{X}\mathbf{U})+\mathrm{tr}(\mathbf{U}^\top\mathbf{X}^\top\mathbf{X}\mathbf{U})\\&+\mathcal{L}_2(\mathbf{U})+n\sum_{\ell\in\mathcal{D}}\sum_{j=1}^p\lambda_{j\ell}|u_{j\ell}|.\end{aligned}$$

Accordingly,

$$\begin{aligned}&\min_{\mathbf{U}}\mathcal{L}(\mathbf{U})-\mathcal{L}(\mathbf{A}_{\mathcal{K}})\\&\geq\min_{\mathbf{U}}\left[\|\mathbf{X}\mathbf{U}-\mathbf{X}\mathbf{A}_{\mathcal{D}}\|^2+\mathrm{tr}\left\{(\mathbf{X}\mathbf{A}_{\mathcal{D}^c})^\top(\mathbf{I}-\mathbf{H}_{\mathbf{U}})\mathbf{X}\mathbf{A}_{\mathcal{D}^c}\right\}\right]\\&-\max_{\mathbf{U}}\left(2\|\mathbf{X}^\top\mathbf{E}_{\mathcal{D}}\|\|\mathbf{A}_{\mathcal{D}}-\mathbf{U}\|+2\|\mathbf{E}_{\mathcal{D}^c}^\top\mathbf{H}_{\mathbf{U}}\mathbf{X}\mathbf{A}_{\mathcal{D}^c}\|+rns_n\gamma_{1n}C\right)\\&-\left\{2\|\mathbf{E}_{\mathcal{D}^c}^\top\mathbf{X}\mathbf{A}_{\mathcal{D}^c}\|+\mathrm{tr}\left(\mathbf{E}_{\mathcal{K}^c}^\top\mathbf{H}_{\mathbf{A}_{\mathcal{K}}}\mathbf{E}_{\mathcal{K}^c}\right)\right\}:=\Delta_1+\Delta_2+\Delta_3,\end{aligned}$$

where $C > 0$ is a constant.

Denote $c_n = \sqrt{nrp \log p}$. Observe that $\|\mathbf{X}^\top \mathbf{E}_{\mathcal{D}}\| = O_p(c_n)$, $\|\mathbf{E}_{\mathcal{D}^c}^\top \mathbf{X}\mathbf{A}_{\mathcal{D}^c}\| = O_p(c_n)$, $\|\mathbf{E}_{\mathcal{D}^c}^\top \mathbf{H}_{\mathbf{U}}\mathbf{X}\mathbf{A}_{\mathcal{D}^c}\| = O_p(c_n)$ and $\mathrm{tr}(\mathbf{E}_{\mathcal{K}^c}^\top \mathbf{H}_{\mathbf{A}_{\mathcal{K}}}\mathbf{E}_{\mathcal{K}^c}) = O_p(rp \log p)$.

By Condition 5, we know that

$$c_n^{-1}\min_{\mathbf{U}}\left[\|\mathbf{X}\mathbf{U} - \mathbf{X}\mathbf{A}_{\mathcal{D}}\|^2 + \mathrm{tr}\left\{(\mathbf{X}\mathbf{A}_{\mathcal{D}^c})^\top(\mathbf{I} - \mathbf{H}_{\mathbf{U}})\mathbf{X}\mathbf{A}_{\mathcal{D}^c}\right\}\right] \to \infty,$$

we have either $c_n^{-1}\|\mathbf{X}\mathbf{U} - \mathbf{X}\mathbf{A}_{\mathcal{D}}\|^2 \to \infty$ or $c_n^{-1}\mathrm{tr}\left\{(\mathbf{X}\mathbf{A}_{\mathcal{D}^c})^\top(\mathbf{I} - \mathbf{H}_{\mathbf{U}})\mathbf{X}\mathbf{A}_{\mathcal{D}^c}\right\} \to \infty$.

Consider the former one. Note that $\|\mathbf{X}\mathbf{U} - \mathbf{X}\mathbf{A}_{\mathcal{D}}\|^2 \lesssim n\bar{\kappa}\|\mathbf{A}_{\mathcal{D}} - \mathbf{U}\|^2$ and thus $\|\mathbf{A}_{\mathcal{D}} - \mathbf{U}\|/\sqrt{c_n/n} \to \infty$. In this case, the $\Delta_1 \geq \min_{\mathbf{U}}\|\mathbf{X}\mathbf{U} - \mathbf{X}\mathbf{A}_{\mathcal{D}}\|^2$ which dominates $\Delta_2$ and $\Delta_3$. Under the situation that the later one holds, it can be similarly shown that the $\Delta_1$ will dominate the other terms. □

**Proof of Theorem 1**

Theorem 1 follows immediately from Lemmas 3-4. □

**Proof of Theorem 2**

Consider $k < r$ firstly. Using the same arguments in the proof of Lemma 4, it can be seen that

$$\min_k \mathrm{BIC}(k) - \mathrm{BIC}(r) \gtrsim \sqrt{nrp \log p} - \sqrt{rp \log p}h_n.$$

It follows immediately that $\Pr(\min_k \mathrm{BIC}(k) > \mathrm{BIC}(r)) \to 1$ as $n \to \infty$, provided that $h_n/\sqrt{n} \to 0$.

For the case $k > r$, we firstly notice that using the same procedure in the proof of Lemma 1, it can be shown that $\|\hat{\mathbf{A}}(k) - \mathbf{A}\| = O_p(\beta_{n,k})$. Accordingly,

$$\begin{aligned}\min_k \mathrm{BIC}(k) - \mathrm{BIC}(r) &\gtrsim O_p(\sqrt{kp \log p}) + (\sqrt{k} - \sqrt{k-1})\sqrt{p \log p}h_n \\ &\geq O_p(\sqrt{kp \log p}) + \frac{1}{2}k^{-1/2}\sqrt{p \log p}h_n,\end{aligned}$$

which implies that with probability tending to one the case of $k > r$ would not happen as long as $k/h_n \to 0$.

Combining the two cases together implies that any $k$ failing to identify the true low-rank structure cannot be selected as the optimal rank. That is to say, the model associated with the optimal $k$ must be the true one. This completes the proof. □

**Acknowledgments**

The authors are grateful to the referees for their insightful comments that have significantly improved the article.

## References

Ahn, S. C. and Horenstein, A. R. (2013), "Eigenvalue ratio test for the number of factors," *Econometrica*, 81, 1203–1227.

Anderson, T. W. (1951), "Estimating linear restrictions on regression coefficients for multivariate normal distributions," *The Annals of Mathematical Statistics*, 22, 327–351.

— (2004), *An Introduction to Multivariate Statistical Analysis, 3rd ed.*, New York: John Wiley and Sons.

Bing, X. and Wegkamp, M. H. (2019), "Adaptive estimation of the rank of the coefficient matrix in high-dimensional multivariate response regression models," *The Annals of Statistics*, 47, 3157–3184.

Bunea, F., She, Y., and Wegkamp, M. H. (2011), "Optimal selection of reduced rank estimators of high-dimensional matrices," *The Annals of Statistics*, 39, 1282–1309.

Fan, J. and Peng, H. (2004), "Nonconcave penalized likelihood with a diverging number of parameters," *The Annals of Statistics*, 32, 928–961.

Franck, U., Odeh, S., Wiedensohler, A., Wehner, B., and Herbarth, O. (2011), "The effect of particle size on cardiovascular disorders–The smaller the worse," *Science of the Total Environment*, 409, 4217–4221.

Kong, Y., Li, D., Fan, Y., and Lv, J. (2017), "Interaction pursuit in high-dimensional multi-response regression via distance correlation," *The Annals of Statistics*, 45, 897–922.

Negahban, S. and Wainwright, M. J. (2011), "Estimation of (near) low-rank matrices with noise and high-dimensional scaling," *The Annals of Statistics*, 1069–1097.

Obozinski, G., Wainwright, M. J., and Jordan, M. I. (2011), "Support union recovery in high-dimensional multivariate regression," *The Annals of Statistics*, 39, 1–47.

Polichetti, G., Cocco, S., Spinali, A., Trimarco, V., and Nunziata, A. (2009), "Effects of particulate matter (PM10, PM2. 5 and PM1) on the cardiovascular system," *Toxicology*, 261, 1–8.

Pun, V. C., Kazemiparkouhi, F., Manjourides, J., and Suh, H. H. (2017), "Long-term PM2. 5 exposure and respiratory, cancer, and cardiovascular mortality in older US adults," *American journal of epidemiology*, 186, 961–969.

Raskutti, G., Yuan, M., and Chen, H. (2019), "Convex regularization for high-dimensional multiresponse tensor regression," *The Annals of Statistics*, 47, 1554–1584.

Reinsel, G. C. and Velu, R. P. (1998), *Multivariate Reduced-Rank Regression Theory and Applications*, Springer.

Riediker, M., Cascio, W. E., Griggs, T. R., Herbst, M. C., Bromberg, P. A., Neas, L., Williams, R. W., and Devlin, R. B. (2004), "Particulate matter exposure in cars is associated with cardiovascular effects in healthy young men," *American journal of respiratory and critical care medicine*, 169, 934–940.

Xing, Y.-F., Xu, Y.-H., Shi, M.-H., and Lian, Y.-X. (2016), "The impact of PM2. 5 on the human respiratory system," *Journal of thoracic disease*, 8, E69.

Yuan, M., Ekici, A., Lu, Z., and Monteiro, R. (2007), "Dimension reduction and coefficient estimation in multivariate linear regression," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69, 329–346.

Zhang, P. (1993), "Model selection via multifold cross validation," *The Annals of Statistics*, 299–313.

Zheng, Z., Bahadori, M. T., Liu, Y., and Lv, J. (2019), "Scalable Interpretable Multi-Response Regression via SEED," *Journal of Machine Learning Research*, 20, 1–34.

Zou, H. (2006), "The adaptive lasso and its oracle properties," *Journal of the American statistical association*, 101, 1418–1429.