ORIGINAL ARTICLE

Journal Section

High-Dimensional Multivariate Linear Regression with Weighted Nuclear Norm Regularization

Namjoon Suh[1] | Li-Hsiang Lin[2] | Xiaoming Huo[1]

[1]H. Milton Stewart School of Industrial Systems and Engineering, Georgia Institute of Technology, Atlanta, Georgia, 30332, USA

[2]Department of Experimental Statistics, Louisiana State University, Baton Rouge, Louisiana, 70803, USA

Correspondence
Author One PhD, Department, Institution, Atlanta, Georgia, 30332, USA
Email: correspondingauthor@email.com

Funding information
Funder One, Funder One Department, Grant/Award Number: 123456, 123457 and 123458; Funder Two, Funder Two Department, Grant/Award Number: 123459

We consider a low-rank matrix estimation problem when the data is assumed to be generated from the multivariate linear regression model. To induce the low-rank coefficient matrix, we employ the weighted nuclear norm (WNN) defined as the weighted sum of singular values. The weights are naturally set in the non-decreasing order, and this order is known to yield the non-convexity of the WNN function in the parameter space. We provide an efficient algorithm under the framework of alternative directional method of multipliers (ADMM) for estimating the coefficient matrix. The estimator from suggested algorithm converges to a stationary point of augmented Lagrangian function. Under orthogonal design setting, effects of the weights for estimating singular values of ground-truth coefficient matrix are given. Under Gaussian design setting, a minimax convergence rate on the estimation error is derived. We also propose a generalized cross-validation (GCV) criterion for the selection of the tuning parameter, and suggest an iterative algorithm for updating the weights. Simulations and a real data analysis demonstrate competitive performance of our new method.

KEYWORDS

Weighted Nuclear Norm, Low-rank Matrix, Non-convex optimization, Generalized Cross Validation

1 | INTRODUCTION

We consider the problem of recovering an unknown coefficient matrix $\mathbf{\Theta}^\star \in \boldsymbol{R}^{d_1 \times d_2}$ from n observations of the response vector $y_i \in \mathbb{R}^{d_2}$ and predictor $x_i \in \mathbb{R}^{d_1}$, where the ground truth model is as follows:

$$\boldsymbol{Y} = \boldsymbol{X}\mathbf{\Theta}^\star + \boldsymbol{E}. \tag{1}$$

Here, $\boldsymbol{Y} = (y_1, \ldots, y_n)^\top$ is an $n \times d_2$ matrix, $\boldsymbol{X} = (x_1, \ldots, x_n)^\top$ is an $n \times d_1$ matrix, and $\boldsymbol{E} = (e_1, \ldots, e_n)^\top$ is an $n \times d_2$ regression noise matrix. The vectors $\{e_j\}_{j=1}^n$ are independently sampled from $\mathcal{N}(0, \sigma^2 \cdot \mathcal{I}_{d_2 \times d_2})$ with variance parameter $\sigma^2 > 0$. Throughout the paper, we write $p := \min(d_1, d_2)$, $r^\star := \mathrm{rank}(\mathbf{\Theta}^\star)$ and $\mathcal{I}_{m \times m}$ as an $m \times m$ identity matrix.

The observational model (1) is referred to as a multivariate linear regression model in the statistics literature. This model is particularly attractive when there exists a dependence structure in the multivariate response, where the response matrix $\boldsymbol{Y}$ can be represented with a linear combination of only a small number of linearly transformed predictors. The situation is induced from the assumption that the coefficient matrix $\mathbf{\Theta}^\star$ has a low rank, that is $r^\star \ll p$.

Given the noisy measurement pair $(\boldsymbol{X}, \boldsymbol{Y})$, estimating the ground-truth $\mathbf{\Theta}^\star$ with the consistent rank has been intensively studied by many researchers during the past decades. Among them, Yuan et al. (2007) suggested the least-square problem with nuclear norm (also known as trace norm) penalization, giving the simultaneous dimension reduction and estimation of the coefficient matrix. Analogous to the use of ℓ_1-regularizer for enforcing sparsity of signal in linear regression setting, nuclear norm is mathematically defined as the sum of singular values of a matrix, and enforces the sparsity in the vector of singular values. However, estimator from the standard nuclear norm (SNN) penalized least-square method still suffers from the bias introduced from the penalization and generally has a higher rank estimate than other methods. In order to mitigate this issue, Chen et al. (2013) employed the idea of weighted nuclear norm (WNN) penalization. The core idea of WNN is to put the small weights on large singular values to reduce the bias and to put the large weights on small singular values to encourage the estimated matrix to have a low rank. Nonetheless, Chen et al. (2013) considered the WNN penalization on $\boldsymbol{X}\mathbf{\Theta}$ instead solely on $\mathbf{\Theta}$, where $\mathbf{\Theta}$ is a parameter of interests for inference.

Along this line of research, we consider the statistical estimation problem with WNN penalization only on the coefficient matrix $\mathbf{\Theta}$ by solving a following optimization problem:

$$\widehat{\mathbf{\Theta}} := \underset{\mathbf{\Theta} \in \mathbb{R}^{d_1 \times d_2}}{\mathrm{argmin}} \left\{ \frac{1}{2n} \|\boldsymbol{Y} - \boldsymbol{X}\mathbf{\Theta}\|_{\mathrm{F}}^2 + \boldsymbol{\lambda}_n \|\mathbf{\Theta}\|_{\boldsymbol{\omega},\star} \right\} \tag{2}$$

with

$$\|\mathbf{\Theta}\|_{\boldsymbol{\omega},\star} = \sum_{j=1}^{p} \omega_j \sigma_j(\mathbf{\Theta}), \tag{3}$$

where $\sigma_j(\mathbf{\Theta})$ means the j^{th} largest singular value of a matrix $\mathbf{\Theta} \in \boldsymbol{R}^{d_1 \times d_2}, \boldsymbol{\omega} = (\omega_1, ..., \omega_p)$, ω_j is a non-negative weight assigned to $\sigma_j(\mathbf{\Theta})$, $\boldsymbol{\lambda}_n \geq 0$ is a hyper-tuning parameter, and $\|\cdot\|_{\mathrm{F}} := \sqrt{\sum_{j=1}^{p} \sigma_j(\cdot)^2}$ is a Frobenius norm. Specifically, it is a well-known fact that the landscape of (2) is non-convex when the weights are in non-decreasing order, $0 \leq \omega_1 \leq \omega_2 \leq \cdots \leq \omega_p$. Hereafter, our paper only considers the case of non-decreasing weights. See the simple example provided in Chen et al. (2013) which shows that (3) is neither convex nor concave function when the weights are in the non-decreasing order. Under this setting, we briefly summarize the contributions of our paper.

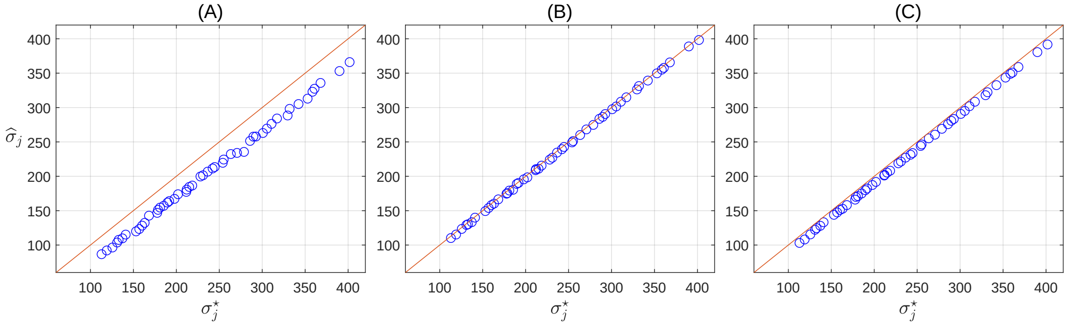

FIGURE 1 Three panels display the plots of estimated sigular values versus ground truth singular values $\sigma_j^\star$. The first two panels (A) and (B) are results from WMVR-ADMM algorithm with one weight update iteration under $n = 250$. The panel (C) exhibits the result when the estimator is obtained from SNN penalized least square under $n = 1000$.

1.1 | Contributions

We apply the classical alternative direction method of multipliers (ADMM) algorithm (Boyd et al., 2011) in (2) and show that the sequence of tuples generated from the suggested algorithm converges uniquely to a stationary point of augmented Lagrangain function. We refer our algorithm as WMVR-ADMM where the WMVR stands for weighted multivariate regression. This should be contrasted with the result from Chen et al. (2013), in which they provide the closed form solution of the $\widehat{\mathbf{\Theta}}$ of (2), not with the penalization $||\mathbf{\Theta}||_{\omega,\star}$ but with $||\boldsymbol{X}\mathbf{\Theta}||_{\omega,\star}$. See Corollary 1 in their paper. Furthermore, the theoretical analysis of Chen et al. (2013) is focused on the behavior of prediction error, not the estimation error which is one of the theoretical objects of interest in our paper.

Our paper provides a theoretical explanation on the role of weights for estimating the ground-truth coefficient matrix. Motivated from Yuan et al. (2007), under the orthogonal design setting, we derive the closed-form solution of the minimizer $\widehat{\mathbf{\Theta}}$ and provide a non-asymptotic convergence rate of its singular values to its ground-truth counterparts. We show that the smaller weights compared to 2σ are desirable for estimating the non-zero $\sigma_j(\mathbf{\Theta}^\star)$s, whereas the larger weights than 2σ are required for estimating zero $\sigma_j(\mathbf{\Theta}^\star)$s under a proper choice of $\boldsymbol{\lambda}_n$. Under a Gaussian random design setting, we derive the minimax rate of the estimation error by adopting the technique used by Negahban and Wainwright (2011) under high-dimensional regime (i.e., $n \ll d_1 d_2$). We will elaborate their work in the following sections in the context of our work.

Finally, we develop a data-driven method for choosing tuning parameters in the model. For updating the weights on the singular value, we borrow the idea from the seminal work of Candes et al. (2008). The algorithm we propose consists of solving a sequence of WNN problems, where the weights used for the next iteration are computed from the singular values of the current solution from (2). Regarding a choice of hyper-tuning parameter $\boldsymbol{\lambda}_n$, we adopt a generalized cross-validation (GCV) type of criterion. This is enabled through the development of a surrogate function (equation (21)), whose solution can be closely approximated to the solution of (2). The solution also allows us to approximate the degrees of freedom of the original multivariate linear regression problem, and this set-up makes the GCV statistic computable.

A following example demonstrates the sample efficiency of our proposed method for estimating the singular values of $\mathbf{\Theta}^\star$ when it is compared with the traditional SNN method. We consider a setting of coefficient matrix $\mathbf{\Theta}^\star \in \mathbb{R}^{250\times250}$ with $r^\star = 50$ and generate $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{250\times50}$ with each entry from $\mathcal{N}(0,1)$ and set $\mathbf{\Theta}^\star = \boldsymbol{A}\boldsymbol{B}^\top$. Each entry of $\boldsymbol{X} \in \mathbb{R}^{n \times d_1}$ is sampled from $\mathcal{N}(0, 1)$. Variance parameter σ is set as 1, and hyper-tuning parameter $\boldsymbol{\lambda}_n$ is set as $5\sqrt{\frac{d_1+d_2}{n}}$, where $d_1 = d_2 = 250$. Panels in Figure 1 display the plots of singular values of the minimizer $\widehat{\boldsymbol{\Theta}}$ in (2) against the singular values of ground-truth matrix $\boldsymbol{\Theta}^{\star}$. Panel (A) exhibits the result of the first iteration of WMVR-ADMM with sample size $n = 250$, and panel (B) exhibits the result of the second iteration of the algorithm with the updated weights. Here, we set the j^{th} weight ω_j as the inverse of the j^{th} singular value of $\widehat{\boldsymbol{\Theta}}$ from the first iteration. Note that we start the WMVR-ADMM with $\{\omega_j\}_{j=1}^{p} = 1$, equivalent to solving SNN problem. Panel (C) displays the result of SNN problem with $n = 1000$. It is worth noting that WMVR-ADMM achieves a satisfactory result within two iterations of loop with only $n = 250$, whereas there is still a slight bias on each of the estimated singular value from SNN with $n = 1000$.

1.2 | Additional Related Literature

In the field of computer vision, many papers including Gu et al. (2014, 2017); Xu et al. (2017); Yair and Michaeli (2018); Liu et al. (2018); Kim et al. (2020) studied WNN minimization problem in the context of matrix completion problem. However, from the statistical viewpoint, we are not aware of many works that apply the WNN in matrix regression problem except Chen et al. (2013).

In contrast, there are a myriad of papers which studied the statistical properties of SNN penalized least square problem under even a more general model than multivariate linear regression. We only mention a subset of the long list. Bach (2008) provided necessary and sufficient conditions for the asymptotic rank consistency of SNN problem, and later Lee et al. (2015) worked on proving the non-asymptotic rank consistency of the estimator from SNN under the irrepresentable assumption on design matrix. Under the sub-Gaussian noise assumption, Negahban and Wainwright (2011) derived a minimax optimal rate of the estimation error of trace regression model when $\boldsymbol{\Theta}^{\star}$ is either approximately or exactly low rank matrix through the employment of the notion of restricted strong convexity (RSC) of the cost function. Similarly, Koltchinskii et al. (2011) established a sharp oracle inequality of the trace regression estimator under the restricted isometry condition of design matrix $\boldsymbol{X}$. In the subsequent work, Fan et al. (2019) investigated the SNN problem under generalized trace regression problems for categorical responses. Recently, Fan et al. (2021) worked on obtaining the same minimax rate of trace regression problem with Negahban and Wainwright (2011) under the heavy-tail assumption on design matrix and observational noise.

Our work also falls into the category of adaptive penalized estimation problem. Among a plethora of papers, we consider that the most relevant work with our paper is Zou (2006), which proposed the adaptive lasso in the context of sparse linear regression. However, it is worth noting that once the weights are fixed, minimizing the least square fit with the adaptive ℓ_1-penalization is always a convex optimization problem. Later, Candes et al. (2008) suggested an algorithm for updating the weights in the adaptive lasso problem. The main idea of their paper is to simply update the weights as the inverse of the estimated signal in the previous iteration.

1.3 | Organizations

The rest of the paper is organized as follows. In Section 2, we introduce the details of WMVR-ADMM and provide a theorem on the algorithm's convergence guarantees. In Section 3, statistical properties of the estimator are provided. First, in the orthogonal design setting, the non-asymptotic convergence rate of the singular values from the proposed estimator $\{\sigma_j(\widehat{\boldsymbol{\Theta}})\}_{j=1}^{p}$ is provided. Second, under a Gaussian random design, we obtain the minimax rate of the estimation error. In Section 4, a two-stage data-driven method for updating weights and choosing the regularization parameter is detailed. In Section 5, we compare the performance of our estimator with SNN and an estimator from Chen et al. (2013) in terms of estimation error under various model parameter settings. In Section 6, we apply our algorithm to a real data set to demonstrate the validity of WMVR-ADMM in practice. Finally, we conclude our paper with the discussion section.

2 | WMVR-ADMM AND CONVERGENCE ANALYSIS

In the first subsection, we describe the implementation of WMVR-ADMM algorithm for solving the non-convex optimization problem (2). Then, in the next subsection, we present a convergence theorem of WMVR-ADMM algorithm.

2.1 | ADMM for Weighted Multivariate Regression

We start with reformulating (2) as follows:

$$\min_{\Theta,\Gamma}\Big\{f(\mathbf{\Theta})+g(\mathbf{\Gamma})\Big\} \qquad \textbf{s.t.} \qquad \mathbf{\Theta}=\mathbf{\Gamma}\in\mathbb{R}^{d_1\times d_2}, \tag{4}$$

by letting $f(\mathbf{\Theta}) := \boldsymbol{\lambda}_n||\mathbf{\Theta}||_{\omega,\star}$ and $g(\mathbf{\Gamma}) = \frac{1}{2n}\|\boldsymbol{Y}-\boldsymbol{X}\mathbf{\Gamma}\|_{\mathrm{F}}^2$. This reformulation naturally leads to the construction of an augmented lagrangian function $\mathcal{L}_\rho(\mathbf{\Theta},\mathbf{\Gamma},\mathbf{\Lambda})$: For any $\rho>0$ and dual variable $\mathbf{\Lambda}\in\mathbb{R}^{d_1\times d_2}$, we define,

$$\mathcal{L}_\rho(\mathbf{\Theta},\mathbf{\Gamma},\mathbf{\Lambda}) := f(\mathbf{\Theta})+g(\mathbf{\Gamma})+\mathbf{tr}\big(\mathbf{\Lambda}^\top(\mathbf{\Theta}-\mathbf{\Gamma})\big)+\frac{\rho}{2}||\mathbf{\Theta}-\mathbf{\Gamma}||_{\mathrm{F}}^2. \tag{5}$$

Then, we solve following three optimization problems repeatedly until primal and dual feasibility condition hold; that is, repeat **Steps 1-3**,

$$\begin{aligned}
&\textbf{Step 1.} \quad \mathbf{\Theta}^{(k+1)} = \underset{\Theta\in\mathbb{R}^{d_1\times d_2}}{\text{argmin}}\ \mathcal{L}_\rho(\mathbf{\Theta},\mathbf{\Gamma}^{(k)},\mathbf{\Lambda}^{(k)}),\\
&\textbf{Step 2.} \quad \mathbf{\Gamma}^{(k+1)} = \underset{\Gamma\in\mathbb{R}^{d_1\times d_2}}{\text{argmin}}\ \mathcal{L}_\rho(\mathbf{\Theta}^{(k+1)},\mathbf{\Gamma},\mathbf{\Lambda}^{(k)}),\\
&\textbf{Step 3.} \quad \mathbf{\Lambda}^{(k+1)} = \mathbf{\Lambda}^{(k)}+\rho\big(\mathbf{\Theta}^{(k+1)}-\mathbf{\Gamma}^{(k+1)}\big),
\end{aligned}$$

until $||\mathbf{\Theta}^{(k+1)}-\mathbf{\Gamma}^{(k+1)}||_{\mathrm{F}}\le 10^{-7}$ and $||\mathbf{\Gamma}^{(k+1)}-\mathbf{\Gamma}^{(k)}||_{\mathrm{F}}\le 10^{-7}$. Here, we denote the tuple $(\mathbf{\Theta}^{(k)},\mathbf{\Gamma}^{(k)},\mathbf{\Lambda}^{(k)})$ as the updated parameters at k^{th} iteration of the algorithm. Note that the non-convexity of the landscape of the objective function in **Step 1** arises from the WNN (i.e., $\|\cdot\|_{\omega,\star}$) over $\mathbf{\Theta}$ with fixed $\mathbf{\Gamma}^{(k)},\mathbf{\Lambda}^{(k)}$, whereas the objective function in **Step 2** is a simple quadratic function of $\mathbf{\Gamma}$ with fixed $\mathbf{\Theta}^{(k+1)},\mathbf{\Lambda}^{(k)}$.

We start the algorithm by initializing $\mathbf{\Theta}^{(0)}=\mathbf{\Gamma}^{(0)}=\mathbf{\Lambda}^{(0)}=\mathbf{0}\in\mathbb{R}^{d_1\times d_2}$. Next, the key of our algorithm is that a closed-form solution of **Step 1** can be obtained uniquely, even if it is a non-convex problem. We state the result in the following Lemma.

Lemma 1 *Let $\mathbf{\Theta}^{(k+1)}$ be the minimizer of **Step 1**. Denote $\boldsymbol{B}^{(k)} := -\mathbf{\Lambda}^{(k)}+\rho\cdot\mathbf{\Gamma}^{(k)}$ and its SVD as $\boldsymbol{U}^{\mathbf{B}}\boldsymbol{D}^{\mathbf{B}}(\boldsymbol{V}^{\mathbf{B}})^\top$. Then, for any fixed $\boldsymbol{\lambda}_n,\rho\ge 0$ and $0\le\omega_1\le\cdots\le\omega_p$,*

$$\begin{aligned}
&\mathbf{\Theta}^{(k+1)} = \boldsymbol{U}^{\mathbf{B}}\mathcal{S}_{\boldsymbol{\lambda}_n\omega}(\boldsymbol{D}^{\mathbf{B}})(\boldsymbol{V}^{\mathbf{B}})^\top,\\
&\mathcal{S}_{\boldsymbol{\lambda}_n\omega}(\boldsymbol{D}^{\mathbf{B}}) = diag\Big\{\max\Big\{\frac{1}{\rho}\big(\sigma_j(\boldsymbol{B}^{(k)})-\boldsymbol{\lambda}_n w_j\big),0\Big\}, j=1,\ldots,p\Big\}.
\end{aligned}$$

Furthermore, if all the non-zero singular values of $\boldsymbol{B}^{(k)}$ are distinct, then the solution $\mathbf{\Theta}^{(k+1)}$ is unique.

Proof For simplicity, denote $\boldsymbol{B}^{(k)} := -\mathbf{\Lambda}^{(k)} + \rho \cdot \mathbf{\Gamma}^{(k)}$, then we can solve the optimization problem in **Step 1** as follows:

$$\begin{aligned} \mathbf{\Theta}^{(k+1)} &= \underset{\mathbf{\Theta} \in \mathbb{R}^{d_1 \times d_2}}{\operatorname{argmin}} \ \mathcal{L}_\rho(\mathbf{\Theta}, \mathbf{\Gamma}^{(k)}, \mathbf{\Lambda}^{(k)}) \\ &= \underset{\mathbf{\Theta} \in \mathbb{R}^{d_1 \times d_2}}{\operatorname{argmin}} \left\{ f(\mathbf{\Theta}) + \mathbf{tr}\big(\mathbf{\Lambda}^{(k)\top}\mathbf{\Theta}\big) + \frac{\rho}{2}||\mathbf{\Theta} - \mathbf{\Gamma}^{(k)}||_{\mathrm{F}}^2 \right\} \\ &= \underset{\mathbf{\Theta} \in \mathbb{R}^{d_1 \times d_2}}{\operatorname{argmin}} \left\{ \sum_{j=1}^{p} \left(\frac{\rho}{2}\sigma_j(\mathbf{\Theta})^2 + \boldsymbol{\lambda}_n \omega_j \cdot \sigma_j(\mathbf{\Theta}) \right) - \mathbf{tr}\big(\boldsymbol{B}^{(k)\top}\mathbf{\Theta}\big) \right\}. \end{aligned} \tag{6}$$

We plugged-in $f(\mathbf{\Theta}) = \boldsymbol{\lambda}_n||\mathbf{\Theta}||_{\boldsymbol{\omega},\star}$, used $\mathbf{tr}(\mathbf{\Theta}\mathbf{\Theta}^\top) = \sum_{j=1}^{p} \sigma_j(\mathbf{\Theta})^2$ and the definition of $\boldsymbol{B}^{(k)}$ for deriving the last equality. For further convenience of notation, let $\{d_j\}_{j=1}^p := \{\sigma_j(\mathbf{\Theta})\}_{j=1}^p$ and denote $\mathbf{\Theta} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top$ where $\boldsymbol{U}$ and $\boldsymbol{V}$ are the left and right singular matrices of $\mathbf{\Theta}$ and $\boldsymbol{D} := \operatorname{diag}(\{d_1, d_2, \ldots, d_p\})$. Note that the entries in $\boldsymbol{D}$ are in non-increasing order. (i.e. $d_1 \geq d_2 \cdots \geq d_p \geq 0$) Then, we can rewrite the optimization problem in (6) as follows:

$$\mathbf{\Theta}^{(k+1)} = \underset{d_1 \geq d_2 \geq \cdots \geq d_p \geq 0}{\operatorname{argmin}} \left\{ \sum_{j=1}^{p} \left(\frac{\rho}{2}d_j^2 + \boldsymbol{\lambda}_n \omega_j d_j \right) - \max_{\boldsymbol{U}^\top\boldsymbol{U}=\mathcal{I}_{d_1}, \boldsymbol{V}^\top\boldsymbol{V}=\mathcal{I}_{d_2}} \mathbf{tr}\big(\boldsymbol{B}^{(k)\top}\mathbf{\Theta}\big) \right\} \tag{7}$$

The maximum of second term in (7) can be achieved when $\boldsymbol{U}$ and $\boldsymbol{V}$ coincide with left and right singular matrices of $\boldsymbol{B}^{(k)}$ respectively, giving us the maximized value as $\sum_{j=1}^{p} \sigma_j(\boldsymbol{B}^{(k)})d_j$. This is a well-known Von Neumann's trace inequality. See Von Neumann (1937); Mirsky (1975). Then, the final form of the optimization problem (7) reduces to obtaining the diagonal entries of the matrix $\boldsymbol{D}$ by minimizing a following :

$$\min_{d_1 \geq d_2 \geq \cdots \geq d_p \geq 0} \left\{ \sum_{j=1}^{p} \left(\frac{\rho}{2}d_j^2 + \big(\boldsymbol{\lambda}_n\omega_j - \sigma_j(\boldsymbol{B}^{(k)})\big)d_j \right) \right\}. \tag{8}$$

The objective function (8) is completely decompsable coordinate-wise and is minimized at $d_j = \max\left\{\frac{1}{\rho}\big(\sigma_j(\boldsymbol{B}^{(k)}) - \boldsymbol{\lambda}_n\omega_j\big), 0\right\}$ for $j = 1, \ldots, p$. Since $\sigma_1(\boldsymbol{B}^{(k)}) \geq \sigma_2(\boldsymbol{B}^{(k)}) \cdots \geq \sigma_p(\boldsymbol{B}^{(k)})$ and $0 \leq \omega_1 \leq \omega_2 \leq \cdots \leq \omega_p$, the solution is feasible. Furthermore, we have an unique minimizer due to the equality condition of von-Neumann's trace inequality when $\boldsymbol{B}^{(k)}$ has distinct non-zero singular values, and the uniqueness of strict convex optimization of (8) in d_j for $j = 1, \ldots, p$. □

Now, we turn our attention on minimizing the convex program in **Step 2**. It is easy to rewrite and solve the problem in **Step 2** as follows:

$$\begin{aligned} \mathbf{\Gamma}^{(k+1)} &= \underset{\mathbf{\Gamma} \in \mathbb{R}^{d_1 \times d_2}}{\operatorname{argmin}} \ \mathcal{L}_\rho(\mathbf{\Theta}^{(k+1)}, \mathbf{\Gamma}, \mathbf{\Lambda}^{(k)}) \\ &= \underset{\mathbf{\Gamma} \in \mathbb{R}^{d_1 \times d_2}}{\operatorname{argmin}} \left\{ \mathbf{tr}\left(\mathbf{\Gamma}^\top \left(\frac{1}{2n}\boldsymbol{X}^\top\boldsymbol{X} + \frac{\rho}{2} \cdot \mathcal{I}_{d_1 \times d_1} \right) \mathbf{\Gamma} - \left(\frac{1}{n}\boldsymbol{Y}^\top\boldsymbol{X} + \rho \cdot \mathbf{\Theta}^{(k+1)} + \mathbf{\Lambda}^{(k)} \right)^\top \mathbf{\Gamma} \right) \right\} \end{aligned} \tag{9}$$

$$= \left(\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X} + \rho \cdot \mathcal{I}_{d_1 \times d_1} \right)^{-1} \left(\frac{1}{n}\boldsymbol{Y}^\top\boldsymbol{X} + \rho \cdot \mathbf{\Theta}^{(k+1)} + \mathbf{\Lambda}^{(k)} \right). \tag{10}$$

Note that the quadratic equation (9) always has an unique minimizer (10) as long as $\rho > 0$. With the updated $\mathbf{\Theta}^{(k+1)}$ and $\mathbf{\Gamma}^{(k+1)}$ from **Steps 1**, **2**, we can easily update $\mathbf{\Lambda}^{(k)}$ to $\mathbf{\Lambda}^{(k+1)}$ through **Step 3**. The final output of WMVR-ADMM is a minimizer of $\mathcal{L}_\rho(\mathbf{\Theta}, \mathbf{\Gamma}^{(\mathcal{T}-1)}, \mathbf{\Lambda}^{(\mathcal{T}-1)})$ in **Step 1**, where $\mathcal{T}$ denotes the last iteration index of the algorithm. Therefore,

Algorithm 1: ADMM for Weighted Multi-Variate Regression. (WMVR-ADMM)

```
Input : A measurement pair $(\boldsymbol{X}, \boldsymbol{Y})$, $\boldsymbol{\lambda}_n \geq 0$ and $0 \leq \omega_1 \leq \cdots \leq \omega_p$.
Initialization : $\mathbf{\Theta}^{(0)} = \mathbf{\Gamma}^{(0)} = \mathbf{\Lambda}^{(0)} = \mathbf{0} \in \mathbb{R}^{d_1 \times d_2}$.
Repeat following Steps :
    Step 1. Let $\boldsymbol{B}^{(k)} := -\mathbf{\Lambda}^{(k)} + \rho \cdot \mathbf{\Gamma}^{(k)}$.    $\boldsymbol{B}^{(k)} = \boldsymbol{U}^{\mathbf{B}} \boldsymbol{D}^{\mathbf{B}} (\boldsymbol{V}^{\mathbf{B}})^{\top}$.
            Set $\mathcal{S}_{\boldsymbol{\lambda}_n \omega}(\boldsymbol{D}^{\mathbf{B}}) = \textbf{diag}\Big\{ \max\Big\{ \frac{1}{\rho}(\sigma_j(\boldsymbol{B}^{(k)}) - \boldsymbol{\lambda}_n w_j), 0 \Big\}$ for $j = 1, \ldots, p \Big\}$.
            $\mathbf{\Theta}^{(k+1)} = \boldsymbol{U}^{\mathbf{B}} \mathcal{S}_{\boldsymbol{\lambda}_n \omega}(\boldsymbol{D}^{\mathbf{B}}) (\boldsymbol{V}^{\mathbf{B}})^{\top}$
    Step 2. $\mathbf{\Gamma}^{(k+1)} = \big(\frac{1}{n}\boldsymbol{X}^{\top}\boldsymbol{X} + \rho \cdot \mathcal{I}_{d_1 \times d_1}\big)^{-1} \big(\frac{1}{n}\boldsymbol{Y}^{\top}\boldsymbol{X} + \rho \cdot \mathbf{\Theta}^{(k+1)} + \mathbf{\Lambda}^{(k)}\big)$.
    Step 3. $\mathbf{\Lambda}^{(k+1)} = \mathbf{\Lambda}^{(k)} + \rho\big(\mathbf{\Theta}^{(k+1)} - \mathbf{\Gamma}^{(k+1)}\big)$.
Until $||\mathbf{\Theta}^{(k+1)} - \mathbf{\Gamma}^{(k+1)}||_{\mathrm{F}} \leq 10^{-7}$ and $||\mathbf{\Gamma}^{(k+1)} - \mathbf{\Gamma}^{(k)}||_{\mathrm{F}} \leq 10^{-7}$.
Output : $\widehat{\mathbf{\Theta}} = \mathbf{\Theta}^{(k+1)}$.
```

as long as all the non-zero singular values of $\boldsymbol{B}^{(\mathcal{T})}$ are distinct, then the WMVR-ADMM has an unique solution. The entire implementation of WMVR-ADMM is summarized in Algorithm 1.

Remark We want to note that WMVR-ADMM algorithm can be easily extended to trace regression model, which is a general model of multivariate linear regression model.[1] In order for the concise presentation of the paper, we defer the detailed descriptions of the extended algorithm to trace regression model in the Appendix 8.1.

2.2 | Convergence of WMVR-ADMM

In this subsection, the Theorem 2 on the convergence of WMVR-ADMM is presented with some remarks. Then, two lemmas for proof of the theorems are provided. Finally, we conclude the subsection with the proof of Theorem 2.

Theorem 2 *Set $\rho > 2L_{\nabla g}$ with $L_{\nabla g} := \sigma_1\big(\frac{1}{n}\boldsymbol{X}^{\top}\boldsymbol{X}\big)$. The sequence $\{(\mathbf{\Theta}^{(k)}, \mathbf{\Gamma}^{(k)}, \mathbf{\Lambda}^{(k)})\}_{k \geq 1}$ from WMVR-ADMM converges globally to the unique stationary point of $\mathcal{L}_{\rho}(\mathbf{\Theta}, \mathbf{\Gamma}, \mathbf{\Lambda})$.*

The threshold for penalty parameter ρ can be computed from data. Here, "globally" means regardless of where the initial point under non-convex landscape of (5). This is verified numerically in subsection 5.1. The uniqueness claim is due from the fact that the augmented lagrangian function (5) is a Kurdyka-Lojasiewicz (KL) function. This is further elaborated in sub-subsection 2.2.1 with relevant references. The proof of Theorem 2 is motivated from Wang et al. (2019) and Kim et al. (2020). The key two ingredients are provided as follows. The detailed proofs of following two Lemmas are in the Appendix.

Lemma 3 *Set $\rho > 2L_{\nabla g}$ with $L_{\nabla g} := \sigma_1\big(\frac{1}{n}\boldsymbol{X}^{\top}\boldsymbol{X}\big)$. Then, the iterates $\{(\mathbf{\Theta}^{(k)}, \mathbf{\Gamma}^{(k)}, \mathbf{\Lambda}^{(k)})\}_{k \geq 1}$ generated from WMVR-ADMM satisfy followings :*

1. *$\mathcal{L}_{\rho}\big(\mathbf{\Theta}^{(k)}, \mathbf{\Gamma}^{(k)}, \mathbf{\Lambda}^{(k)}\big)$ is lower-bounded and non-increasing over $k \geq 1$.*
2. *$\{(\mathbf{\Theta}^{(k)}, \mathbf{\Gamma}^{(k)}, \mathbf{\Lambda}^{(k)})\}_{k \geq 1}$ is bounded.*
3. *$\Big\|\mathbf{\Theta}^{(k)} - \mathbf{\Gamma}^{(k)}\Big\|_F \to 0$ and $\Big\|\mathbf{\Gamma}^{(k+1)} - \mathbf{\Gamma}^{(k)}\Big\|_F \to 0$, as $k \to \infty$.*

Lemma 4 *For $k \geq 1$, there exist a constant $C_2 > 0$ and $p^{(k+1)} \in \partial \mathcal{L}_{\rho}\big(\mathbf{\Theta}^{(k+1)}, \mathbf{\Gamma}^{(k+1)}, \mathbf{\Lambda}^{(k+1)}\big)$ such that $\|p^{(k+1)}\|_F \leq C_2 \|\mathbf{\Gamma}^{(k+1)} - \mathbf{\Gamma}^{(k)}\|_F$.*

[1] Refer Negahban and Wainwright (2011) for checking how to translate MVLR to trace regression model.

2.2.1 | Proof of Theorem 2

The proof of Theorem 2 is divided into five steps. We provide each steps sequentially.

Step 1. By Bolzano-Weierstrass threorem, we know the bounded sequence $\{\boldsymbol{\Theta}^{(k)}, \boldsymbol{\Gamma}^{(k)}, \boldsymbol{\Lambda}^{(k)}\}_{k\geq 0}$ has a convergent subsequence $\{\boldsymbol{\Theta}^{(k_s)}, \boldsymbol{\Gamma}^{(k_s)}, \boldsymbol{\Lambda}^{(k_s)}\}_{s\geq 1}$, and denote its limit point as $(\boldsymbol{\Theta}^{*}, \boldsymbol{\Gamma}^{*}, \boldsymbol{\Lambda}^{*})$.

Step 2. Since the augmented lagrangian function $\mathcal{L}_{\rho}(\boldsymbol{\Theta}^{(k)}, \boldsymbol{\Gamma}^{(k)}, \boldsymbol{\Lambda}^{(k)})$ is non-increasing and bounded from below, the sequence $\{\mathcal{L}_{\rho}(\boldsymbol{\Theta}^{(k)}, \boldsymbol{\Gamma}^{(k)}, \boldsymbol{\Lambda}^{(k)})\}_{k\geq 0}$ converges.

Step 3. By continuity of $\mathcal{L}_{\rho}(\boldsymbol{\Theta}^{(k)}, \boldsymbol{\Gamma}^{(k)}, \boldsymbol{\Lambda}^{(k)})$ and results from **Step 1** and **Step 2**, we have

$$\lim_{k\to\infty} \mathcal{L}_{\rho}(\boldsymbol{\Theta}^{(k)}, \boldsymbol{\Gamma}^{(k)}, \boldsymbol{\Lambda}^{(k)}) = \lim_{s\to\infty} \mathcal{L}_{\rho}(\boldsymbol{\Theta}^{(k_s)}, \boldsymbol{\Gamma}^{(k_s)}, \boldsymbol{\Lambda}^{(k_s)}) = \mathcal{L}_{\rho}(\boldsymbol{\Theta}^{*}, \boldsymbol{\Gamma}^{*}, \boldsymbol{\Lambda}^{*}).$$

Step 4. By the result of Lemma 2.4, there exists $p^{(k+1)} \in \partial\mathcal{L}_{\rho}(\boldsymbol{\Theta}^{(k+1)}, \boldsymbol{\Gamma}^{(k+1)}, \boldsymbol{\Lambda}^{(k+1)})$ such that $\|p^{(k+1)}\|_{\mathrm{F}} \to 0$ as $k \to \infty$. Consequently, we conclude the following:

$$p^{(k+1)} \in \partial\mathcal{L}_{\rho}(\boldsymbol{\Theta}^{(k+1)}, \boldsymbol{\Gamma}^{(k+1)}, \boldsymbol{\Lambda}^{(k+1)}) \to \mathbf{0} \in \partial\mathcal{L}_{\rho}(\boldsymbol{\Theta}^{*}, \boldsymbol{\Gamma}^{*}, \boldsymbol{\Lambda}^{*}), \quad k \to \infty.$$

Step 5. It remains to prove $\mathcal{L}_{\rho}$ is a Kurdyka-Lojasiewicz (KL) function for ensuring the generated sequence $\{(\boldsymbol{\Theta}^{(k)}, \boldsymbol{\Gamma}^{(k)}, \boldsymbol{\Lambda}^{(k)})\}_{k\geq 0}$ converges globally to the unique point $\{\boldsymbol{\Theta}^{*}, \boldsymbol{\Gamma}^{*}, \boldsymbol{\Lambda}^{*}\}$. See Proposition 2 in Wang et al. (2019) and Theorem 2.9 in Attouch et al. (2013). If a function is semi-algebraic, then it is known to be a KL function (Attouch et al., 2013; Sun et al., 2017). Refer Sun et al. (2017) for the definition of semi-algebraic function. Since $\frac{1}{2n}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\Gamma}\|_{\mathrm{F}}^{2} + \mathbf{tr}(\boldsymbol{\Lambda}^{\top}(\boldsymbol{\Theta} - \boldsymbol{\Gamma}))$ is a real-polynomial function, it is semi-algebraic. Since the finite sum of semi-algebraic functions are semi-algebraic (Attouch et al., 2013), it remains to prove $\lambda_n\|\boldsymbol{\Theta}\|_{\omega,*}$ is a semi-algebraic. In proposition 3 of Sun et al. (2017), it is proved that each singular value of the matrix $\boldsymbol{\Theta}$, $\sigma_j(\boldsymbol{\Theta})$, is a semi-algebraic. Therefore, the weighted singular value $\omega_j\sigma_j(\boldsymbol{\Theta})$ is also a semi-algebraic function, and finally, summing rule gives that $\lambda_n\|\boldsymbol{\Theta}\|_{\omega,*}$ is a semi-algebraic function. □

3 | STATISTICAL PROPERTIES OF THE ESTIMATOR

In this section, we explore statistical properties of the proposed estimator. In subsection 3.1, the non-asymptotic property of the proposed estimator under orthogonal design setting is studied, which sheds lights on understanding the role of weights on the estimation of singular values. In subsection 3.2, the minimax convergence rate of the estimation error is derived under a Gaussian design setting.

3.1 | Statistical Properties of $\widehat{\boldsymbol{\Theta}}$ under Orthogonal Design

We first study the convergent rate of estimated eigenvalues $\sigma_j(\widehat{\boldsymbol{\Theta}})$ for $j = 1, \cdots, p$. The result with its proof is provided below.

Proposition 5 *Suppose . Let $\widehat{\boldsymbol{U}}^{LS}\widehat{\boldsymbol{D}}^{LS}(\widehat{\boldsymbol{V}}^{LS})^{\top}$ be SVD of the least-square estimator $\widehat{\boldsymbol{\Theta}}^{LS} := (\boldsymbol{X}^{\top}\boldsymbol{X})^{-1}\boldsymbol{X}^{\top}\boldsymbol{Y}$. Then, under the orthogonal design (i.e., $\boldsymbol{X}^{\top}\boldsymbol{X} = n\boldsymbol{I}_{d_1\times d_1}$), SVD of the minimizer of* (2) *has a following closed-form solution: $\widehat{\boldsymbol{\Theta}} :=$*

$\widehat{\boldsymbol{U}}^{\mathrm{LS}}\widehat{\boldsymbol{D}}(\widehat{\boldsymbol{V}}^{\mathrm{LS}})^\top$, where the diagonal entry of $\widehat{\boldsymbol{D}}$ is: $\sigma_j(\widehat{\boldsymbol{\Theta}}) = \max\left(\sigma_j(\widehat{\boldsymbol{\Theta}}^{LS}) - \boldsymbol{\lambda}_n\omega_j, 0\right)$ for $j = 1, \dots, p$. Furthermore, suppose $\boldsymbol{\lambda}_n = \sqrt{\frac{d_1+d_2}{n}}$. Then, with probability at least $1 - 2\exp(-(\sqrt{d_1}+\sqrt{d_2})^2/2)$, we have,

$$\left|\sigma_j(\widehat{\boldsymbol{\Theta}}) - \sigma_j(\boldsymbol{\Theta}^\star)\right| \le \max\left(4\sigma, 2\omega_j\right) \cdot \sqrt{\frac{d_1+d_2}{n}}, \tag{11}$$

for j such that $\sigma_j(\boldsymbol{\Theta}^\star) > 0$. With the same probability bound, we have,

$$\left|\sigma_j(\widehat{\boldsymbol{\Theta}})\right| \le \min\left(2\sigma, \omega_j\right) \cdot \sqrt{\frac{d_1+d_2}{n}}, \tag{12}$$

for j such that $\sigma_j(\boldsymbol{\Theta}^\star) = 0$.

Proof The derivation on the closed-form solution of $\widehat{\boldsymbol{\Theta}}$ is exactly same with that of Lemma 1 in Yuan et al. (2007), under the orthogonal design assumption. So we omit the proof. We only focus on controlling the distance between singular values of $\widehat{\boldsymbol{\Theta}}$ and $\boldsymbol{\Theta}^\star$. With the equality $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\Theta}^\star + \boldsymbol{E}$ and $\boldsymbol{X}^\top\boldsymbol{X} = n\boldsymbol{I}_{d_1\times d_1}$, we have

$$\widehat{\boldsymbol{\Theta}}^{\mathrm{LS}} = \left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top\boldsymbol{Y} = \boldsymbol{\Theta}^\star + \frac{\boldsymbol{X}^\top\boldsymbol{E}}{n}. \tag{13}$$

By the corollary of Weyl's Theorem and the equality (13), inequality

$$\max_{j=1,\dots,p}\left|\sigma_j(\widehat{\boldsymbol{\Theta}}^{\mathrm{LS}}) - \sigma_j(\boldsymbol{\Theta}^\star)\right| \le \sigma_1\left(\frac{\boldsymbol{X}^\top\boldsymbol{E}}{n}\right). \tag{14}$$

can be obtained. Recall from our problem setting that the rows of $\boldsymbol{E}$ are independent from $\mathcal{N}(0, \sigma^2\boldsymbol{I}_{d_2\times d_2})$. Therefore, we know each entry of $\boldsymbol{X}^\top\boldsymbol{E}/\sigma\sqrt{n}$ follows $\mathcal{N}(0,1)$ and is independent with each other. Following Chapter 6 in Wainwright (2019), with probability at least $1 - 2\exp(-(\sqrt{d_1}+\sqrt{d_2})^2/2)$, the right hand side of (14) satisfies

$$\sigma_1\left(\frac{\boldsymbol{X}^\top\boldsymbol{E}}{n}\right) \le 2\sigma\sqrt{\frac{d_1+d_2}{n}}. \tag{15}$$

Because $\sigma_j(\widehat{\boldsymbol{\Theta}}) = \sigma_j(\widehat{\boldsymbol{\Theta}}^{\mathrm{LS}}) - \boldsymbol{\lambda}_n w_j > 0$ for $j = 1, \dots, \widehat{r}$, with further combing (14) and (15), we know for $j \in \{1, \dots, \widehat{r}\}$

$$\begin{aligned}
\left|\sigma_j(\widehat{\boldsymbol{\Theta}}) - \sigma_j(\boldsymbol{\Theta}^\star)\right| &= \left|\sigma_j(\widehat{\boldsymbol{\Theta}}^{\mathrm{LS}}) - \boldsymbol{\lambda}_n\omega_j - \sigma_j(\boldsymbol{\Theta}^\star)\right| \\
&\le \left|\sigma_j(\widehat{\boldsymbol{\Theta}}^{\mathrm{LS}}) - \sigma_j(\boldsymbol{\Theta}^\star)\right| + \boldsymbol{\lambda}_n\omega_j \\
&\le \sigma_1\left(\frac{\boldsymbol{X}^\top\boldsymbol{E}}{n}\right) + \boldsymbol{\lambda}_n\omega_j \le \max\left(4\sigma, 2\omega_j\right) \cdot \sqrt{\frac{d_1+d_2}{n}},
\end{aligned}$$

where in the last inequality, we use (15) and choose $\boldsymbol{\lambda}_n = \sqrt{\frac{d_1+d_2}{n}}$. For $j \in \{\widehat{r}+1, \dots, p\}$ such that $\sigma_j(\boldsymbol{\Theta}^\star) > 0$, the following inequalities hold

$$\begin{aligned}
\left|\sigma_j(\widehat{\boldsymbol{\Theta}}) - \sigma_j(\boldsymbol{\Theta}^\star)\right| &\le \left|\sigma_j(\widehat{\boldsymbol{\Theta}}^{\mathrm{LS}}) - \sigma_j(\boldsymbol{\Theta}^\star)\right| + \left|\sigma_j(\widehat{\boldsymbol{\Theta}}^{\mathrm{LS}})\right| \\
&\le \sigma_1\left(\frac{\boldsymbol{X}^\top\boldsymbol{E}}{n}\right) + \boldsymbol{\lambda}_n\omega_j \le \max\left(4\sigma, 2\omega_j\right) \cdot \sqrt{\frac{d_1+d_2}{n}},
\end{aligned}$$

where in the second inequality, we use (14) and $|\sigma_j(\widehat{\mathbf{\Theta}}^{\mathrm{LS}})| \leq \boldsymbol{\lambda}_n \omega_j$ for $j \in \{\widehat{r}+1, \ldots, p\}$. For $j \in \{\widehat{r}+1, \ldots, p\}$ such that $\sigma_j(\mathbf{\Theta}^\star) = 0$, we have the following result:

$$\left|\sigma_j(\widehat{\mathbf{\Theta}}) - \sigma_j(\mathbf{\Theta}^\star)\right| \leq \left|\sigma_j(\widehat{\mathbf{\Theta}}^{\mathrm{LS}})\right| \leq \boldsymbol{\lambda}_n \omega_j. \tag{16}$$

Using that the three inequalities (14), (15), and (16) should hold at the same time, we can conclude the proof. □

Based on the closed-form solution of $\widehat{\mathbf{\Theta}}$ in Proposition 5, under the orthogonal design assumption, each estimated singular value has a form $\max\left(\sigma_j(\widehat{\mathbf{\Theta}}^{\mathrm{LS}}) - \boldsymbol{\lambda}_n \omega_j, 0\right)$ for $j \in \{1, \ldots, p\}$. Then, for the fixed $\boldsymbol{\lambda}_n$, it is easy to see that the large weights for small singular values of $\widehat{\mathbf{\Theta}}^{\mathrm{LS}}$ can induce the sparsity among the singular values of $\widehat{\mathbf{\Theta}}$. Furthermore, the proposition states that with an appropriate choice of tuning parameter $\boldsymbol{\lambda}_n$, the singular values of the $\widehat{\mathbf{\Theta}}$ are consistently estimated. Bounds in (11) and (12) provide us with the guideline for the choices of weights. That is, for the set of indices of $\sigma_j(\mathbf{\Theta}^\star) > 0$, the corresponding weights ω_js need to be set lower than the twice of variance size of the measurement error σ, whereas, for the set of indices whose $\sigma_j(\mathbf{\Theta}^\star) = 0$, the corresponding weights can be set even higher than 2σ. This is consistent with our intuition that we need small weights for estimating non-zero singular values of $\mathbf{\Theta}^\star$, whereas large weights are required for the consistent estimation of zero singular values of $\mathbf{\Theta}^\star$.

3.2 | Estimation Error under Random Design

We further study the estimation error in Frobenius norm (i.e., $\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^\star\|_{\mathrm{F}}^2$) under a random design assumption. Specifically, we begin this subsection by providing the additional assumptions on model, which will be used for stating our theorem. Then, with a brief introduction of the framework on which our theorem is based, an important Lemma on the subset $\mathcal{C}$, where the error matrix $\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^\star$ belongs, is provided. Finally, we state our main Theorem.

3.2.1 | Additional Assumptions

A design matrix $\boldsymbol{X}$ is assumed to be random, whose rows are independently sampled from d_1-variate $\mathcal{N}(0, \mathbf{\Sigma})$ distribution for some positive definite covariance matrix $\mathbf{\Sigma} \in \mathbb{R}^{d_1 \times d_1}$. Additionally, we relax the assumption of $\mathbf{\Theta}^\star$ from a exact low rank setting to a nearly low-rank matrix by requiring that the $\{\sigma_j(\mathbf{\Theta}^\star)\}_{j=1}^p$ decays fast enough. Specifically, for a parameter $q \in [0, 1]$ and a radius $r^\star$, we assume that

$$\mathbf{\Theta}^\star \in \mathbb{B}_q(r^\star) := \left\{\mathbf{\Theta} \in \mathbb{R}^{d_1 \times d_2} : \sum_{j=1}^{p} |\sigma_j(\mathbf{\Theta}^\star)|^q \leq r^\star\right\}.$$

Note that when $q = 0$, the set $\mathbb{B}_q(r^\star)$ becomes the set of matrices with rank at most $r^\star$. Additionally, we assume that $\widehat{\mathbf{\Theta}}$ is a global minimizer of (2).

3.2.2 | General Framework and Characterization of Subset $\mathcal{C}$

There are two key ingredients that will be used for studying the estimation errors : (I) restricted strong convexity of the cost function $\mathcal{L}_n(\mathbf{\Theta}) := \frac{1}{2n}\|\boldsymbol{Y} - \boldsymbol{X}\mathbf{\Theta}\|_{\mathrm{F}}^2$ around $\Theta^\star$ and (II) the characterization of set where the associated error matrix $\widehat{\mathbf{\Delta}} = \widehat{\mathbf{\Theta}} - \mathbf{\Theta}^\star$ belongs. In high-dimensional setting where $n \ll d_1 d_2$, although the function $\mathcal{L}_n(\mathbf{\Theta})$ might be curved in some directions, there are $(d_1 d_2 - n)$ directions where it is flat up to second order. We hope that the associated error matrix $\widehat{\mathbf{\Delta}}$ lies in some directions $\mathcal{C} \subseteq \mathbb{R}^{d_1 \times d_2}$ where the $\mathcal{L}_n(\mathbf{\Theta})$ is curved. This notion is neatly expressed as follows: for some positive constant $\kappa > 0$,

$$\mathcal{E}_n(\widehat{\boldsymbol{\Delta}}) \geq \kappa ||\widehat{\boldsymbol{\Delta}}||_{\mathrm{F}}^2 \qquad \text{for all} \quad \widehat{\boldsymbol{\Delta}} \in \mathcal{C}, \tag{17}$$

where $\mathcal{E}_n(\widehat{\boldsymbol{\Delta}})$ denotes the first order Taylor-expansion error of $\mathcal{L}_n(\cdot)$ around $\boldsymbol{\Theta}^\star$. In other words, we call $\mathcal{E}_n(\widehat{\boldsymbol{\Delta}})$ succeeds *"restricted strong convexity"* (RSC) over the set $\mathcal{C}$ if there exists $\kappa > 0$. Fortunately, under multivariate regression model with Gaussian ensemble, we can prove that the RSC condition indeed holds with $\kappa = \frac{\sigma_{\min}(\boldsymbol{\Sigma})}{18}$ in high probability over $\mathbb{R}^{d_1 \times d_2}$, [2] where $\sigma_{\min}(\boldsymbol{\Sigma})$ denotes a minimum eigenvalue of $\boldsymbol{\Sigma}$.

Before we formally state the Lemma which characterizes the set $\mathcal{C}$, let us introduce relevant notations. Denote $\boldsymbol{U}^\star$ and $\boldsymbol{V}^\star$ as the left and right singular matrices of $\boldsymbol{\Theta}^\star$. The $\mathcal{M}_r(\mathcal{U}, \mathcal{V})$ (resp. $\overline{\mathcal{M}}_r^\perp(\mathcal{U}, \mathcal{V})$) corresponds to subspace of matrices with non-zero left and right singular vectors associated with the first r (resp. remaining $(p-r)$) columns of $\boldsymbol{U}^\star$ and $\boldsymbol{V}^\star$. That is, for any given integer $r \leq p$, we have

$$\mathcal{M}_r(\mathcal{U}, \mathcal{V}) = \{\boldsymbol{\Theta} \in \mathbb{R}^{d_1 \times d_2} : \mathbf{colspan}(\boldsymbol{\Theta}) \subseteq \mathcal{U}, \quad \mathbf{rowspan}(\boldsymbol{\Theta}) \subseteq \mathcal{V}\}$$
$$\overline{\mathcal{M}}_r^\perp(\mathcal{U}, \mathcal{V}) = \{\boldsymbol{\Theta} \in \mathbb{R}^{d_1 \times d_2} : \mathbf{colspan}(\boldsymbol{\Theta}) \subseteq \mathcal{U}^\perp, \quad \mathbf{rowspan}(\boldsymbol{\Theta}) \subseteq \mathcal{V}^\perp\}.$$

Then, $\mathcal{U}$ and $\mathcal{V}$ are the r-dimensional subspaces of vectors from the first r columns of matrices $\boldsymbol{U}^\star$ and $\boldsymbol{V}^\star$. Moreover, $\mathcal{U}^\perp$ and $\mathcal{V}^\perp$ denote the subspaces orthogonal to $\mathcal{U}$ and $\mathcal{V}$, respectively, and $\mathbf{colspan}(\boldsymbol{\Theta})$ and $\mathbf{rowspan}(\boldsymbol{\Theta})$ denote the column space and row space of $\boldsymbol{\Theta}$. Hereafter, we will omit $\mathcal{U}$ and $\mathcal{V}$ from the notations, if they are clear from the context.

Lemma 6 *Suppose $\widehat{\boldsymbol{\Theta}}$ is an global minimizer of the* (2) *obtained from WMVR-ADMM, with the associated matrix $\widehat{\boldsymbol{\Delta}} = \widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^\star$. Set the weights $\frac{1}{2} < \omega_1 \leq \cdots \leq \omega_p$ and suppose regularization parameter is chosen such that $\boldsymbol{\lambda}_n \geq \frac{2}{n}\left\|\boldsymbol{X}^\top \boldsymbol{E}\right\|_{op}$. Let $\|\cdot\|_\star := \sum_{j=1}^p \sigma_j(\cdot)$. Then, for a positive integer $r \leq p$, we have*

$$\mathcal{C}(\omega; r; \delta) := \left\{\widehat{\boldsymbol{\Delta}} \in \mathbb{R}^{d_1 \times d_2} : ||\widehat{\boldsymbol{\Delta}}''||_\star \leq \frac{2w_p}{w_1 - \frac{1}{2}} \sum_{j=r+1}^{p} \sigma_j(\boldsymbol{\Theta}^\star) + \frac{2w_p - w_1 + \frac{1}{2}}{w_1 - \frac{1}{2}} \cdot \|\widehat{\boldsymbol{\Delta}}'\|_\star\right\}, \tag{18}$$

where $\widehat{\boldsymbol{\Delta}}'' \in \Pi_{\overline{\mathcal{M}}_r^\perp}(\widehat{\boldsymbol{\Delta}})$ and $\widehat{\boldsymbol{\Delta}}' = \widehat{\boldsymbol{\Delta}} - \widehat{\boldsymbol{\Delta}}''$. Let $\Pi_{\overline{\mathcal{M}}_r^\perp}$ denote the projection operator onto the subspace $\overline{\mathcal{M}}_r^\perp$.

Now, we state some remarks on the Lemma.

1. The subset $\mathcal{C}$ corresponds to the matrices $\widehat{\boldsymbol{\Delta}}$ for which the quantity $||\widehat{\boldsymbol{\Delta}}''||_\star$ is relatively small compared to the weighted sum of $||\widehat{\boldsymbol{\Delta}}'||_\star$ and $(p-r)$ remaining singular values of $\boldsymbol{\Theta}^\star$. The weights put in $||\widehat{\boldsymbol{\Delta}}'||_\star$ and $\sum_{j=r+1}^p \sigma_j(\boldsymbol{\Theta}^\star)$ are functions of a pair (ω_1, ω_p), and this pair characterizes size of the subset $\mathcal{C}$. We restrict the case $\omega_1 > \frac{1}{2}$ for a technical reason. The closer ω_1 gets to $\frac{1}{2}$ and the larger ω_p we have, the bigger the size of $\mathcal{C}$ becomes.
2. It is worth noting that plugging in $\omega_1 = \cdots = \omega_p = 1$ recovers one of constraints that are used to define the set in Lemma 1 of Negahban and Wainwright (2011). A notable difference between the set in (18) and the set defined in Negahban and Wainwright (2011) is the existence of the constraint, $||\widehat{\boldsymbol{\Delta}}||_{\mathrm{F}} \geq \delta$, where $\delta > 0$ is a tolerance parameter. This constraint is used to eliminate the open ball that is contained within the set $\mathcal{C}$, to ensure RSC condition holds over $\mathcal{C}$, even when $\mathcal{E}_n(\widehat{\boldsymbol{\Delta}})$ fails strong convexity in a global sense. Nonetheless, as previously

[2] See Lemma 2 in Negahban and Wainwright (2011) and Appendix 8.5 of our paper for the proof on this fact.

mentioned, since strong convexity of $\mathcal{L}_n(\boldsymbol{\Theta})$ holds globally in our problem setting, the constraint $||\widehat{\boldsymbol{\Delta}}||_{\mathrm{F}} \geq \delta$ is not required.

3. A detailed proof of the Lemma is deferred in the Appendix. In this remark, we only provide key ideas for the proof. Given that the global minimizer of (2) is obtainable, basic inequality can be employed at the beginning of the proof, which is a standard technique for obtaining the estimation error in various statistical models. However, a technical difficulty for proving the Lemma 6 arises from the fact that the function $||\cdot||_{w,\star}$ is neither convex nor concave in $0 \leq \omega_1 \leq \cdots \leq \omega_p$. We circumvent this problem by rewriting $||\cdot||_{\omega,\star}$ as the difference of two convex functions: that is, $\|\cdot\|_{\omega,\star} = \omega_p\|\cdot\|_{\star} - \|\cdot\|_{\omega_p-\omega,\star}$, where $||\cdot||_{w_p-w,\star} := \sum_{j=1}^{p}(\omega_p - \omega_j)\sigma_j(\cdot)$. Readers should note that since $\omega_p - \omega_1 \geq \cdots \geq \omega_p - \omega_{p-1} \geq 0$, the $\|\cdot\|_{\omega_p-\omega,\star}$ are convex function over the parameter space. See Theorem 1 of Chen et al. (2013) to check this fact.

With the RSC condition and Lemma 3.2, we can further show in the next subsection that the estimation error converges to 0 in a minimax rate.

3.2.3 | Convergence Rate of Estimation Error

Theorem 7 *Suppose* $\boldsymbol{\Theta}^{\star} \in \mathbb{B}_q(r^{\star})$. *The regularization parameter is chosen such that* $\boldsymbol{\lambda}_n = 10\sigma\|\boldsymbol{\Sigma}\|_{op}\sqrt{\frac{d_1+d_2}{n}}$ *and weights are set as* $\frac{1}{2} < \omega_1 \leq \cdots \leq \omega_p$. *Define* $\mathcal{W} := \frac{w_p\left(2w_p - w_1 + \frac{1}{2}\right)}{w_1 - \frac{1}{2}}$. *Then, there are universal constants* $\{c_i, i = 1, 2, 3\}$ *such that any minimizer* $\widehat{\boldsymbol{\Theta}}$ *of* (2) *satisfies a following bound:*

$$\left\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^{\star}\right\|_F^2 \leq c_1\mathcal{W}^2\left(\frac{\sigma^2\|\boldsymbol{\Sigma}\|_{op}^2}{\sigma_{min}^2(\boldsymbol{\Sigma})}\right)^{1-q/2} \cdot r^{\star}\left(\frac{d_1+d_2}{n}\right)^{1-q/2}. \tag{19}$$

with probability at least $1 - c_2\exp(-c_3(d_1 + d_2))$.

Here, $\|\boldsymbol{\Sigma}\|_{\mathrm{op}}$ denotes the spectral norm of the matrix $\boldsymbol{\Sigma}$. Notably, when $\boldsymbol{\Theta}^{\star} \in \mathbb{B}_q(r^{\star})$ is an exact rank $r^{\star}$ matrix (i.e., $q = 0$) and $\boldsymbol{\Sigma} = \mathcal{I}_{d_1 \times d_1}$, convergence rate of the estimation error becomes $O(\mathcal{W}^2\frac{\sigma^2 r^{\star}(d_1+d_2)}{n})$ up to a constant factor. The quantity $r^{\star}(d_1 + d_2)$ counts the degrees of freedom in the model, and the rate is known to be minimax optimal for estimating a $d_1 \times d_2$ matrix with rank $r^{\star}$. See Negahban and Wainwright (2011); Koltchinskii et al. (2011); Rohde and Tsybakov (2011). It is worth noting that the information on weights is solely encoded in factor $\mathcal{W}$. This factor allows a natural comparison of estimation rates between SNN and WNN, and we defer the discussion on this comparison to Section 7.

4 | DATA-DRIVEN MODEL SELECTIONS

In the first subsection, we introduce a surrogate estimator $\widehat{\boldsymbol{\Theta}}^{\mathrm{SR}}$, which approximates the estimator $\widehat{\boldsymbol{\Theta}}$. It is worth noting that $\widehat{\boldsymbol{\Theta}}^{\mathrm{SR}}$ has a ridge regression type of closed-form solution. In the next subsection, the closed form solution is used to calculate the GCV statistic for choosing the proper hyper-tuning parameter $\boldsymbol{\lambda}_n$. Furthermore, a data-driven method for updating the weights is also provided.

4.1 | Surrogate Estimator $\widehat{\boldsymbol{\Theta}}^{\mathrm{SR}}$ for GCV statistic

From Proposition 5, we know $\widehat{\boldsymbol{\Theta}} := \widehat{\boldsymbol{U}}^{\mathrm{LS}}\widehat{\boldsymbol{D}}(\widehat{\boldsymbol{V}}^{\mathrm{LS}})^{\top}$, where the diagonal entry of $\widehat{\boldsymbol{D}}$ is: $\sigma_j(\widehat{\boldsymbol{\Theta}}) = \max\left(\sigma_j(\widehat{\boldsymbol{\Theta}}^{\mathrm{LS}}) - \boldsymbol{\lambda}_n\omega_j, 0\right)$ for $j \in \{1, \dots, p\}$. Hereafter, for the convenience of notation, we denote $\widehat{d_j} := \sigma_j(\widehat{\boldsymbol{\Theta}})$, for $j \in \{1, \dots, p\}$. Then, we define a following matrix $\boldsymbol{K} \in \mathbb{R}^{d_1 \times d_1}$:

$$\boldsymbol{K} := \widehat{\boldsymbol{U}}^{\mathrm{LS}}\widehat{\boldsymbol{D}}^{\mathrm{K}}(\widehat{\boldsymbol{U}}^{\mathrm{LS}})^{\top} := \sum_{j=1}^{\widehat{r}} \frac{\omega_j}{\widehat{d_j}}\widehat{\boldsymbol{U}}_j^{\mathrm{LS}}(\widehat{\boldsymbol{U}}_j^{\mathrm{LS}})^{\top}, \tag{20}$$

where $\widehat{r}$ denotes the cardinality of a set $\{j : \widehat{d_j} > 0\}$. We provide a following proposition.

Proposition 8 *For a fixed $\boldsymbol{K}$, we denote $\widehat{\boldsymbol{\Theta}}^{SR}$ as the minimizer of a following surrogate optimization problem :*

$$\widehat{\boldsymbol{\Theta}}^{SR} := \underset{\boldsymbol{\Theta} \in \mathbb{R}^{d_1 \times d_2}}{\mathrm{argmin}} \left\{ \frac{1}{2n}\|\boldsymbol{Y} - \boldsymbol{X}\Theta\|_F^2 + \frac{\boldsymbol{\lambda}_n}{2}\boldsymbol{tr}(\boldsymbol{\Theta}^{\top}\boldsymbol{K}\boldsymbol{\Theta}) \right\}. \tag{21}$$

Then, under orthogonal design (i.e., $\boldsymbol{X}^{\top}\boldsymbol{X} = n\boldsymbol{I}_{d_1 \times d_1}$), $\widehat{\boldsymbol{\Theta}}^{SR} = \widehat{\boldsymbol{U}}^{LS}\widehat{\boldsymbol{D}}^{SR}(\widehat{\boldsymbol{V}}^{LS})^{\top}$, where

$$\widehat{\boldsymbol{D}}_{jj}^{SR} = \begin{cases} \widehat{d_j} & j = 1, 2, \dots, \widehat{r}, \\ \sigma_j(\widehat{\boldsymbol{\Theta}}^{LS}) & j = \widehat{r} + 1, \dots, p. \end{cases}$$

Proof Let $\boldsymbol{\Theta} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^{\top}$ be the SVD of $\boldsymbol{\Theta}$. Then, we have

$$\mathbf{tr}(\boldsymbol{\Theta}^{\top}\boldsymbol{K}\boldsymbol{\Theta}) = \mathbf{tr}(\boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^{\top}\boldsymbol{K}\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^{\top}) = \mathbf{tr}(\boldsymbol{D}^2\boldsymbol{U}^{\top}\boldsymbol{K}\boldsymbol{U}) = \mathbf{tr}(\boldsymbol{U}\boldsymbol{D}^2\boldsymbol{U}^{\top}\boldsymbol{K}).$$

Let $\boldsymbol{A} := \boldsymbol{U}\boldsymbol{D}^2\boldsymbol{U}^{\top}$ and $\boldsymbol{B} := \boldsymbol{K}$. As proved in Ruhe (1970), for two positive-semi definite matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, we have

$$\mathbf{tr}(\boldsymbol{A}^{\top}\boldsymbol{B}) \geq \sum_{j=1}^{p} \sigma_j(\boldsymbol{A})\sigma_{p+1-j}(\boldsymbol{B}), \tag{22}$$

where $\sigma_1(\cdot) \geq \sigma_2(\cdot) \geq \cdots \geq \sigma_p(\cdot) \geq 0$. Denote $d_j := \sigma_j(\boldsymbol{\Theta})$ for $j \in \{1, \dots, p\}$. Given $0 \leq \omega_1 \leq \omega_2 \leq \cdots \leq \omega_p$, it is easy to see that

$$\sum_{j=1}^{p} \sigma_j(\boldsymbol{A})\sigma_{p+1-j}(\boldsymbol{B}) = \sum_{j=1}^{\widehat{r}} \frac{\omega_j}{\widehat{d_j}} d_j^2. \tag{23}$$

Recalling the assumption $\boldsymbol{X}^{\top}\boldsymbol{X} = n\boldsymbol{I}_{d_1 \times d_1}$ and a simple fact $(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\Theta}}^{\mathrm{LS}})^{\top}\boldsymbol{X} = 0$, we can rewrite the cost function in (21) as follows:

$$\frac{1}{2n}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\Theta}\|_{\mathrm{F}}^2 = \frac{1}{2n}\mathbf{tr}((\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\Theta}}^{\mathrm{LS}})^{\top}(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\Theta}}^{\mathrm{LS}})) + \frac{1}{2}\mathbf{tr}((\widehat{\boldsymbol{\Theta}}^{\mathrm{LS}} - \boldsymbol{\Theta})^{\top}(\widehat{\boldsymbol{\Theta}}^{\mathrm{LS}} - \boldsymbol{\Theta})). \tag{24}$$

By combining (23) and (24), we can obtain the lower bound of the objective function in (21) as follows:

$$\frac{1}{2n}\|\boldsymbol{Y}-\boldsymbol{X}\Theta\|_{\mathrm{F}}^{2}+\frac{\boldsymbol{\lambda}_n}{2}\mathbf{tr}(\boldsymbol{\Theta}^{\top}\boldsymbol{K}\boldsymbol{\Theta})\geq\frac{1}{2}\sum_{j=1}^{p}d_j^2-\sum_{j=1}^{p}\sigma_j(\widehat{\boldsymbol{\Theta}}^{\mathrm{LS}})d_j+\frac{\boldsymbol{\lambda}_n}{2}\sum_{j=1}^{\widehat{r}}\frac{\omega_j}{\widehat{d}_j}d_j^2.$$

We use the equality $\mathbf{tr}(\boldsymbol{\Theta}^{\top}\boldsymbol{\Theta})=\sum_{j=1}^{p}d_j^2$ and the inequality (22) to get a lower bound. It also should be noted that the equality in the above lower bound holds when $\boldsymbol{U}=\widehat{\boldsymbol{U}}^{\mathrm{LS}}$ and $\boldsymbol{V}=\widehat{\boldsymbol{V}}^{\mathrm{LS}}$. Solving the quadratic equation yields the followings:

$$\widehat{\boldsymbol{D}}_{jj}^{\mathrm{SR}}=\begin{cases}\frac{\widehat{d}_j\sigma_j(\widehat{\boldsymbol{\Theta}}^{\mathrm{LS}})}{\widehat{d}_j+\boldsymbol{\lambda}_n\omega_j} & j=1,\dots,\widehat{r},\\ \sigma_j(\widehat{\boldsymbol{\Theta}}^{\mathrm{LS}}) & j=\widehat{r}+1,\dots,p.\end{cases}\tag{25}$$

Recall that $\widehat{d}_j=\sigma_j(\widehat{\boldsymbol{\Theta}}^{\mathrm{LS}})-\boldsymbol{\lambda}_n\omega_j$ for $j=1,\dots,\widehat{r}$, and plugging this equality in (25) for $j\in\{1,\dots,\widehat{r}\}$ yields the claim. □

Note that as long as $\widehat{\boldsymbol{\Theta}}^{\mathrm{LS}}$ is a full-rank, $\widehat{\boldsymbol{\Theta}}^{\mathrm{SR}}$ is a full-rank matrix whose first $\widehat{r}$ singular values are identical to those of $\widehat{\boldsymbol{\Theta}}$, and remaining $(p-\widehat{r})$ singular values are equal to the corresponding singular values of $\widehat{\boldsymbol{\Theta}}^{\mathrm{LS}}$. Although the result of Proposition 8 is stated under orthogonal design assumption, we also empirically demonstrate that the same results hold under non-orthogonal design in Figure 2. Specifically, under the same experimental setting of Figure 1, $\widehat{\boldsymbol{\Theta}}$ is a minimizer of (2) obtained via WMVR-ADMM with the weight updating scheme, which will be introduced in Section 4. In this experiment, the minimum absolute off-diagonal entry of $\boldsymbol{X}^{\top}\boldsymbol{X}$ is 0.00157, which implies $\boldsymbol{X}$ is a non-orthogonal design. The result in Figure 2 is consistent with the statement in Proposition 8.

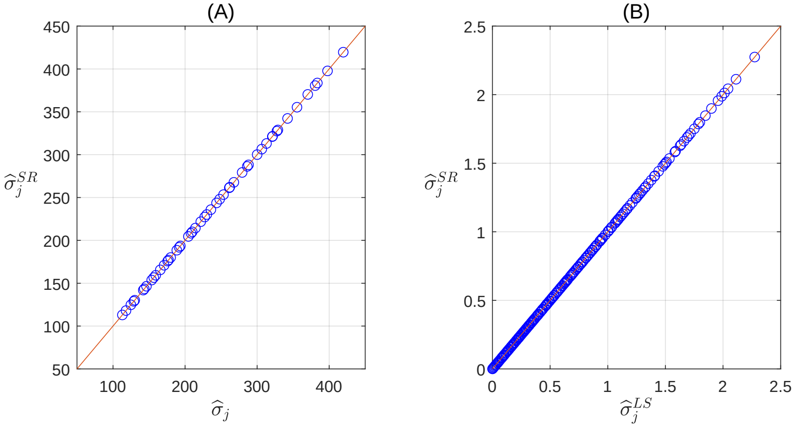

FIGURE 2 Under non-orthogonal $\boldsymbol{X}$, panel (A) displays the plot of the first 50 singular values of $\widehat{\boldsymbol{\Theta}}$ versus $\widehat{\boldsymbol{\Theta}}^{\mathrm{SR}}$. Panel (B) exhibits the plot of the remaining 200 singular values of $\widehat{\boldsymbol{\Theta}}^{\mathrm{LS}}$ versus $\widehat{\boldsymbol{\Theta}}^{\mathrm{SR}}$.

4.2 | GCV Statistic and Weight Updates

We divide the process for tuning $\boldsymbol{\lambda_n}$ and weights $\omega_1, \cdots, \omega_p$ into two procedures. In the first procedure, we propose a following iterative algorithm that alternates between estimating $\boldsymbol{\Theta}^\star$ and updating weights.

(I) Set the iteration count ℓ to 0 and weights $\omega_1^{(0)} = \cdots = \omega_p^{(0)} = 1$.

(II) For the fixed $\boldsymbol{\lambda}_n$, solve (2) via WMVR-ADMM with the weights $\{\omega_j^{(\ell)}\}_{j=1}^p$, and denote the solution as $\widehat{\boldsymbol{\Theta}}^{(\ell)}$.

(III) Update weights : for each $j \in \{1, \ldots, p\}$,

$$\omega_j^{(\ell+1)} = \frac{1}{\sigma_j(\widehat{\boldsymbol{\Theta}}^{(\ell)}) + \epsilon}. \tag{26}$$

(IV) Terminate until convergence or when ℓ attains a pre-specified maximum number of iterations. Otherwise, increment ℓ and go to step **(II)**.

The introduced parameter $\epsilon > 0$ in step **(III)** guarantees that, for any $j \in \{1, \ldots, p\}$, the $(\ell+1)^{\text{th}}$ updated weight $\omega_j^{(\ell+1)}$ is computable, even when $\sigma_j(\widehat{\boldsymbol{\Theta}}^{(\ell)}) = 0$. The recovery process of $\boldsymbol{\Theta}^\star$ is reasonably robust to the choice of ϵ, and we set $\epsilon = 10^{-3}$ hereafter. The choice $\epsilon = 10^{-3}$ may appear a little bit arbitrary, but works well in practice. The resulting estimator from the first procedure is denoted by $\mathrm{W}(\boldsymbol{\lambda_n})$. We use a superscript W in $\widehat{\boldsymbol{\Theta}}^{\mathrm{W}}(\boldsymbol{\lambda_n})$ to indicate that the estimator is a converged solution from weight updating procedure introduced above, and use $\boldsymbol{\lambda_n}$ to denote the estimator is obtained from a fixed hyper-tuning parameter $\boldsymbol{\lambda_n}$.

The second procedure is designed for choosing parameter $\boldsymbol{\lambda_n}$. We develop a GCV type of statistic (Golub et al., 1979), which is more computationally efficient than the ordinary CV (Cross Validation) method, especially in large scale problems. This can be done by using the surrogate estimator $\widehat{\boldsymbol{\Theta}}^{\mathrm{SR}}$ for approximating the degrees of freedom of $\widehat{\boldsymbol{\Theta}}^{\mathrm{W}}(\boldsymbol{\lambda_n})$ from the first procedure. That is, given $\widehat{\boldsymbol{\Theta}}^{\mathrm{W}}(\boldsymbol{\lambda_n})$, we can construct $\boldsymbol{K}^{\mathrm{W}}$ from (20). Then, by proposition 8, we can define the projection matrix (hat matrix) for the regression problem (21) by $\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X} + \boldsymbol{\lambda}_n n\boldsymbol{K}^{\mathrm{W}})^{-1}\boldsymbol{X}^\top$ and approximate the degrees of freedom of $\widehat{\boldsymbol{\Theta}}^{\mathrm{W}}(\boldsymbol{\lambda_n})$ as

$$\mathrm{df}(\boldsymbol{\lambda}_n) \approx d_2\mathbf{tr}(\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X} + \boldsymbol{\lambda}_n n\boldsymbol{K}^{\mathrm{W}})^{-1}\boldsymbol{X}^\top). \tag{27}$$

Thus, the GCV score for $\widehat{\boldsymbol{\Theta}}^{\mathrm{W}}(\lambda_n)$ is given by

$$\mathrm{GCV}(\boldsymbol{\lambda}_n) := \frac{\mathbf{tr}((\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\Theta}}^{\mathrm{W}}(\lambda_n))(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\Theta}}^{\mathrm{W}}(\lambda_n))^\top)}{d_1 d_2 - \mathrm{df}(\boldsymbol{\lambda}_n)}, \tag{28}$$

and the optimal $\boldsymbol{\lambda}_n^\star$ for which GCV($\boldsymbol{\lambda}_n$) is obtained by minimizing the GCV score (28) over the search range $\boldsymbol{\lambda}_n \in [0, \mathcal{T}]$.

5 | SIMULATION STUDY

In subsection 5.1, we empirically demonstrate convergence results stated in Theorem 2. In subsection 5.2, we compare our methods with the existing methods in terms of estimation error. Specifically, SNN method (Yuan et al. (2007)) and the estimator introduced by Chen et al. (2013) are used for the comparisons. Hereafter, we refer the estimator in Chen et al. (2013) as Adpative Nuclear Norm estimator (i.e., ANN estimator). Moreover, we provide an experiment showing the effectiveness of our weight updating scheme (26) in comparison to weight setting in Chen et al. (2013).

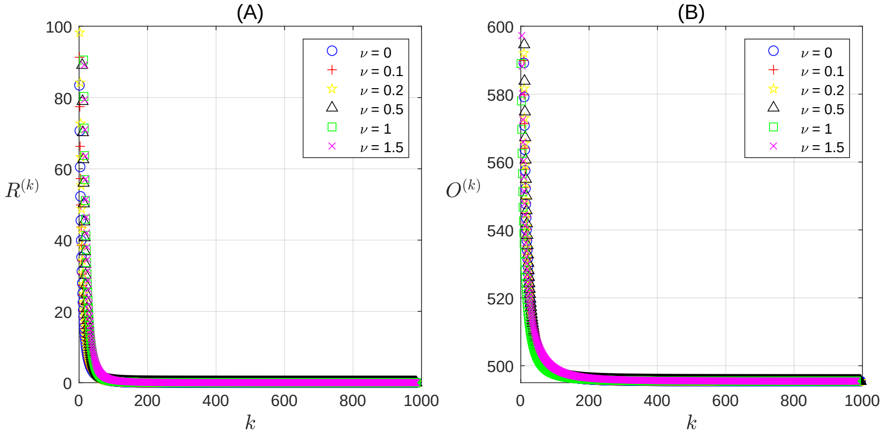

FIGURE 3 Convergences of $R^{(k)}$ (panel (A)) and $O^{(k)}$ (panel (B)) over the algorithm iteration index k. Regardless of random initializations, $R^{(k)}$ and $O^{(k)}$ converge to 0 and to a same objective function value, respectively.

5.1 | Convergence of WMVR-ADMM

The convergences of WMVR-ADMM can be observed through the following two quantities:

1. For checking the Primal residual convergence (i.e., $\mathbf{\Theta}^{(k)} - \mathbf{\Gamma}^{(k)} \to 0$ as $k \to \infty$), and $\mathbf{\Gamma}^{(k)}$ convergence (i.e., $\mathbf{\Gamma}^{(k+1)} - \mathbf{\Gamma}^{(k)} \to 0$ as $k \to \infty$), we consider

$$R^{(k)} := \|\mathbf{\Theta}^{(k)} - \mathbf{\Gamma}^{(k)}\|_F^2 + \|\mathbf{\Gamma}^{(k+1)} - \mathbf{\Gamma}^{(k)}\|_F^2.$$

2. For checking the objective convergence, we consider

$$O^{(k)} := \frac{1}{2n}\left\|\boldsymbol{Y} - \boldsymbol{X}\mathbf{\Theta}^{(k)}\right\|_F^2 + \boldsymbol{\lambda}_n \left\|\mathbf{\Theta}^{(k)}\right\|_{\boldsymbol{\omega},\star}.$$

Under the exactly same simulation setting as that from the experiment for Figure 1 with $n = 250$, we vary the initialized tuple matrix values $(\mathbf{\Theta}^{(0)}, \mathbf{\Gamma}^{(0)}, \mathbf{\Lambda}^{(0)})$ of WMVR-ADMM in Algorithm 1. Entries of three matrices are sampled from $\mathcal{N}(0, \nu^2)$, where $\nu = \{0, 0.1, 0.2, 0.5, 1, 1.5\}$. Weights $\{\omega_j\}_{j=1}^p$ are updated once, and with the updated weights, $R^{(k)}$ and $O^{(k)}$ are calculated with the same data set $(\boldsymbol{X}, \boldsymbol{Y})$ over all simulation scenarios. The resulting $R^{(k)}$ and $O^{(k)}$ values are demonstrated in Figure 3, and the figure shows that both $R^{(k)}$ and $O^{(k)}$ converge to 0 as k increases, regardless of the initializations of algorithm. This observation is consistent with the claims in Theorem 2, and implies that the the converged solutions from WMVR-ADMM have the same objective value on the non-convex landscape of problem (2).

5.2 | Comparisons of Estimation Error with Other Methods

The simulation setting is under model (1) with a ground-truth coefficient matrix $\mathbf{\Theta}^\star$ whose dimension is $d_1 = 25$ and $d_2 = 25$. The coefficient matrix is generated by choosing $\mathbf{\Theta}^\star = \boldsymbol{A}\boldsymbol{B}^\top$, where $\boldsymbol{A}, \boldsymbol{B} \in R^{25 \times r^\star}$ and the elements of $\boldsymbol{A}$

and $\boldsymbol{B}$ are independently and identically following the standardized normal distribution $\mathcal{N}(0,1)$. The value $r^\star$ is the rank of the ground truth matrix and is chosen to be $2, 5, 8$, and 11. The design matrix $\boldsymbol{X}$ are chosen randomly from a normal distribution $\mathcal{N}(0, \mathcal{I}_{d_1\times d_1})$, and the noise matrix $\boldsymbol{E}$ are independently chosen from another $\mathcal{N}(0, \mathcal{I}_{d_2\times d_2})$. The sample sizes n are set to be 30, 300, and 3000, and the simulation is repeated 100 times.

The estimation errors of the proposed method are recorded in terms of the root mean squared errors (RMSE) between the estimated coefficient matrix and the ground-truth matrix for each simulation. The results are compared with those from SNN and ANN methods. Recall SNN estimator is equivalent to the model (1) with $\{\omega_j\}_{j=1}^{p}$. As for ANN estimator, let $\widehat{\boldsymbol{U}}^{\mathrm{XLS}}\widehat{\boldsymbol{D}}^{\mathrm{XLS}}(\widehat{\boldsymbol{V}}^{\mathrm{XLS}})^\top$ be SVD of the matrix $\boldsymbol{X}\widehat{\boldsymbol{\Theta}}^{\mathrm{LS}} := \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{Y}$. Then, the estimator (Corollary 1 in Chen et al. (2013)) has a closed-form solution as:

$$\widehat{\boldsymbol{\Theta}}^{\mathrm{ANN}} = \widehat{\boldsymbol{\Theta}}^{\mathrm{LS}}\widehat{\boldsymbol{V}}^{\mathrm{XLS}}(\widehat{\boldsymbol{D}}^{\mathrm{XLS}})^{-1}\mathcal{S}_{\lambda_n\omega}(\widehat{\boldsymbol{D}}^{\mathrm{XLS}})(\widehat{\boldsymbol{V}}^{\mathrm{XLS}})^\top, \tag{29}$$

where $\mathcal{S}_{\lambda_n\omega}(\widehat{\boldsymbol{D}}^{\mathrm{XLS}}) = \mathrm{diag}\Big\{\max\Big\{\sigma_j(\widehat{\boldsymbol{D}}^{\mathrm{XLS}}) - \lambda_n w_j, 0\Big\} \text{ for } j = 1,\ldots,p\Big\}$. The three methods WMVR-ADMM, SNN, and WNN include parameters needed to be tuned, and we use the GCV tuning method from Section 4 for the tuning.

All results are demonstrated in Figure 5. The first row of Figure 5 shows the performance of all methods under the case whose ground-truth matrix is rank 2 ($r^\star = 2$), and we observe that the averages of RMSEs from the WMVR-ADMM method are smaller than those from other methods for all sample size cases. The second to fourth rows of Figures 5 presents the RMSE results from rank $r^\star = 5, 8$, and 11 cases, and the proposed methods are still better than other methods in almost all cases. Additionally, the panels in Figure 5 demonstrate that the RMSEs from the proposed estimator decrease to 0 as the sample size increases. This shows the consistency property of the proposed estimator empirically.

To show the effectiveness of the proposed weight updating scheme in section 4, we compare the weight setting suggested in Chen et al. (2013) with our method. For the comparison, we revisit the synthetic setting used in Figure 1. Let $\widehat{\boldsymbol{\Theta}}^{(1)}$ be the SNN estimator, and denote ω^{WNN} and ω^{ANN} be the weight settings introduced in (26) and Chen et al.

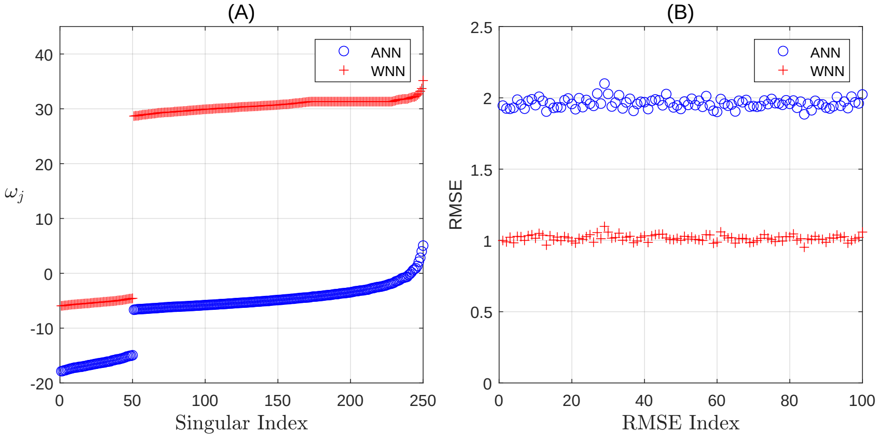

FIGURE 4 Two sequences of weights in (30) used for the estimation (panel (A)) and the resulting RMSEs (panel (B)). Low RMSEs of WNN weights arise from the high penalization on the remaining 200 singular values, when they are compared with RMSEs of ANN weights.

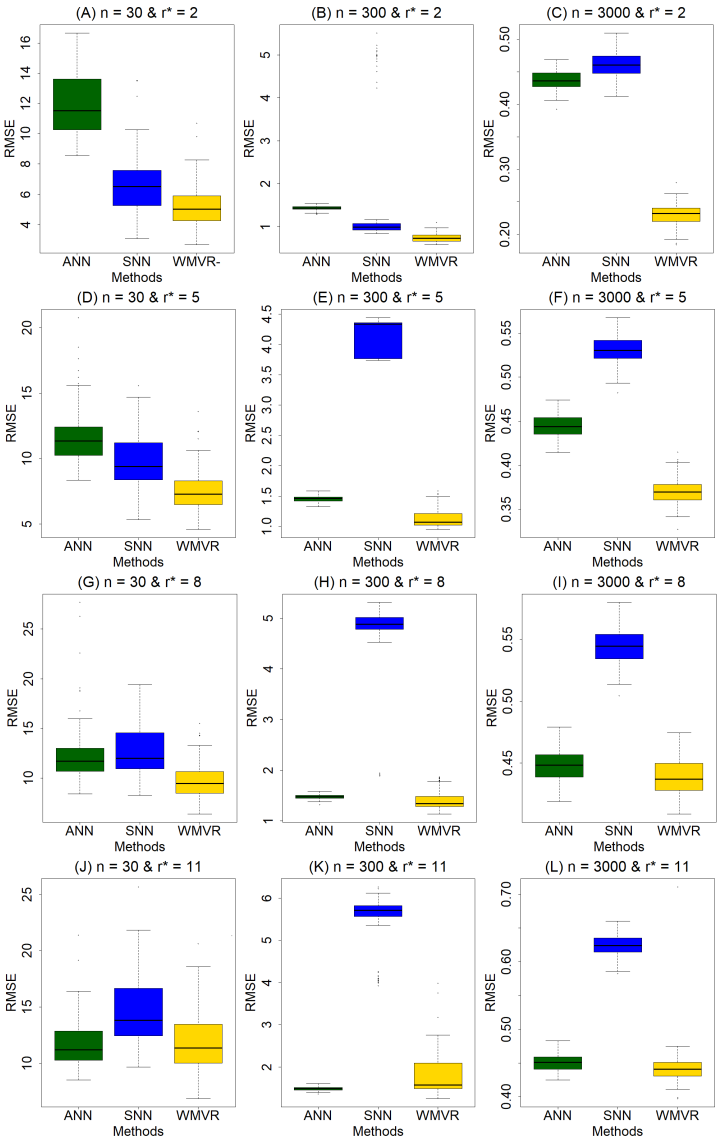

FIGURE 5 The plots demonstrates the comparisons of estimation errors in terms of RMSEs from the proposed method with ANN and SNN methods under different simulation settings. The three figures in the first row (A) ∼ (C) are the comparison results from sample size 30, 300, and 3000, respectively, under the true rank $r^\star = 2$. Analogously, Figures (D) ∼ (F) (second row) are the results from $r^\star = 5$, Figures (G) ∼ (I) (third row) are the results from $r^\star = 8$, and Figures (J) ∼ (L) (fourth row) are the results from $r^\star = 11$.

(2013), respectively. Then, we have

$$\omega_j^{\text{WNN}} = \left(\sigma_j(\widehat{\mathbf{\Theta}}^{(1)}) + 10^{-3}\right)^{-1}, \qquad \omega_j^{\text{ANN}} = \sigma_j(\boldsymbol{X}\widehat{\mathbf{\Theta}}^{\text{LS}})^{-2}, \qquad j = 1, \ldots, 250. \tag{30}$$

For the fair comparison, we use WMVR-ADMM estimator. In Figure 4, panel (A) shows the two sequences of averaged weights $\{\omega_j\}_{j=1}^{250}$ in (30) used for the estimation in logarithmic scale, and panel (B) exhibits 100 RMSEs with the respective weight scheme. While the difference of the first 50 weights between two weight schemes is negligible, the effect of WNN-weight scheme is dramatized for penalizing the remaining 200 singular values in comparison to ANN-weight scheme, and this results in lower RMSEs in panel (B).

6 | APPLICATION TO A REAL DATASET

Although there are many applications of multivariate linear regression (1) in the literature, these applications are usually analyzed by assuming a full rank model and least square method for estimating. In this section, we demonstrate an important application of model (1) without the full rank assumption through the proposed WMVR-ADMM method. The application is about a study of Polycyclic Aromatic Hydrocarbons (PAHs) from Section 2.2.2 of Isenmann (2008).

PAHs are ubiquitous environmental contaminants generated primarily during the incomplete combustion of some organic substances, such as coal, oil, rubbish, and wood. They are linked with the causes of tumors and their effects on reproduction. PAHs are widely used in industry or medicines to make dyes, plastics, and pesticides. In the dataset, 10 PAHs, including pyrene (Py), acenaphthene (Ace), anthracene (Anth), acenaphthylene (Acy), chrysene (Chry), benzanthracene (Benz), fluoranthene (Fluora), fluorene (Fluore), naphthalene (Nap), phenanthracene (Phen), and 25 complex

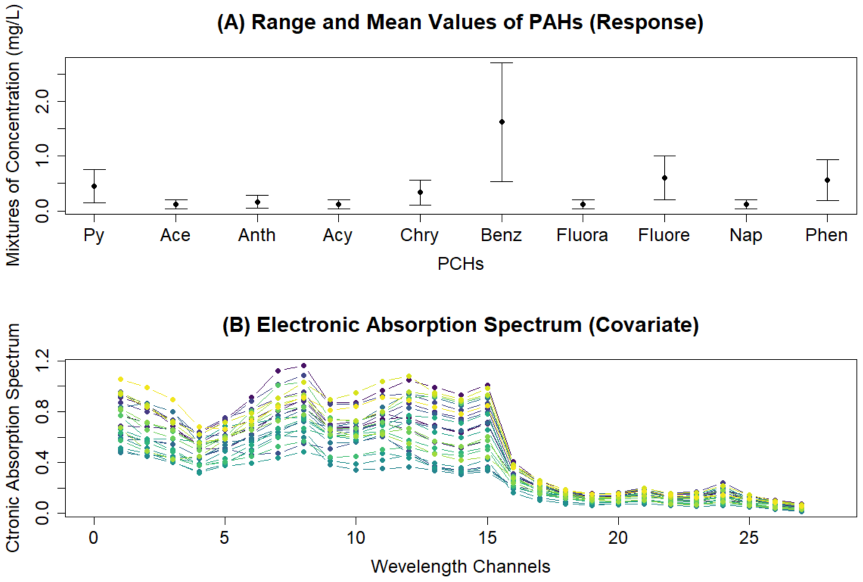

FIGURE 6 Demonstration of the mixture components of the PAHs ($\boldsymbol{Y}$) and the electronic absorption spectrum of the 25 samples ($\boldsymbol{X}$)

mixtures of certain concentrations (with unit milligrams per liter) of these PAHs were recorded, which indicates $n = 25$ and $d_1 = 10$ in model (1). The mean and range values of these mixtures of certain concentrations are plotted in Panel (A) of Figure 6. From each of these mixtures, an electronic absorption spectrum is computed, The spectrum are digitized at 5 nm intervals 27 wavelength channels from 220 nm to 350 nm, as shown in in Panel (B) of Figure 6. This means there are 27 columns for $\boldsymbol{X}_2$ in model (1) ($d_2 = 27$). More details about the dataset can be found in Section 5.1.2 of Brereton (2003) and Section 2.2.2 of Isenmann (2008).

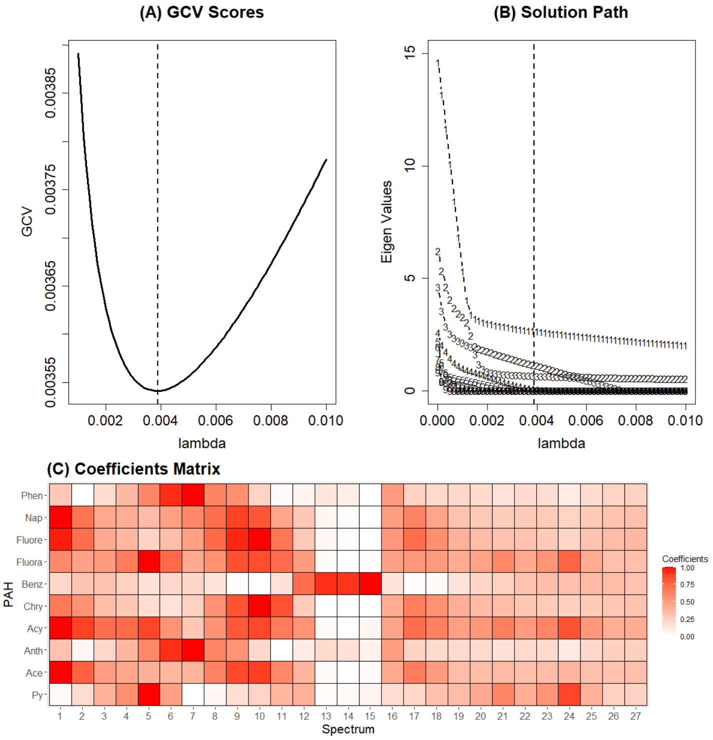

FIGURE 7 (A) GCV Score Versus Tuning Parameters $\boldsymbol{\lambda}$, (B) Solution Path, (C) Estimated Coefficient Matrix.

We are mainly interested in using WMVR-ADMM to understand the association between the concentrations from PAHs (Figure 6 (A)) and the electronic absorption spectrum (Figure 6 (B)) through model (1). The method is conducted by following Algorithm 1, and the optimal tuning parameter $\boldsymbol{\lambda}_n$ and weights $\boldsymbol{w}$ are selected by the proposed GCV criterion described in Section 4. The resulting GCV scores are plotted in Figure 7 (A) with respect to value $\boldsymbol{\lambda}_n$, showing the selected $\boldsymbol{\lambda}_n$ is around 0.039. The estimated eigenvalues with respect to $\boldsymbol{\lambda}_n$ are plotted in Figure 7 (B), and under the optimal $\boldsymbol{\lambda}_n$ and weights from the GCV criterion, the estimated coefficient matrix is rank 5. The estimated coefficients are demonstrated in a heatmap as shown in Figure 7 (C). The figure shows that for each PAH, only a few important channels can be used to determine the concentrations because only some coefficients are relatively large. Additionally, these larger coefficients are usually from smaller column numbers in the heatmap. Thus, the channels with smaller wavelengths are more important than larger wavelength channels.

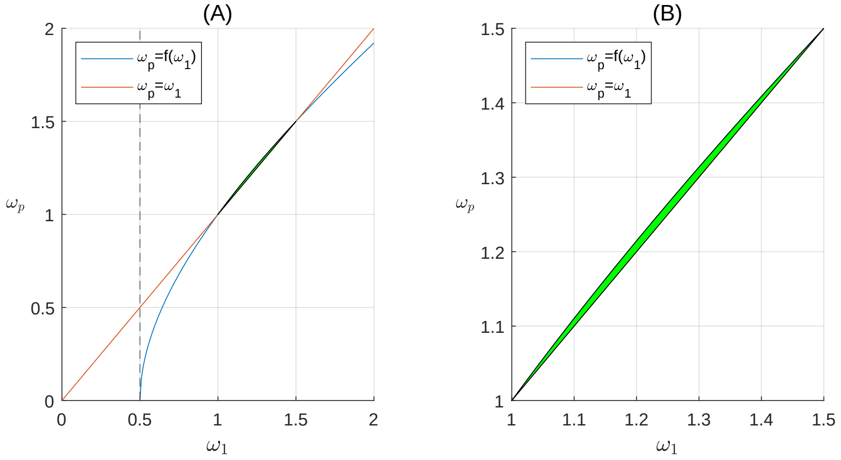

FIGURE 8 Panel (A) exhibits the intersected region of $\mathcal{W} \leq 3$ and $\frac{1}{2} < \omega_1 \leq \cdots \leq \omega_p$. Panel (B) magnifies the intersected region on grid $(\omega_1, \omega_p) \in [1, 1.5] \times [1, 1.5]$.

7 | CONCLUSION AND DISCUSSION

We propose an ADMM-based method for solving the multivariate regression problem with WNN penalty. Under non-decreasing order of weights, the WNN is a non-convex function, and induce non-convexity of WNN penalized least-square problem in (1) over the parameter space. The provided algorithm is shown to converge uniquely to one of stationary points of augmented Lagrangian function. The statistical properties of the estimator are investigated under orthogonal design, providing some insights on the choices of weights for the estimation. Furthermore, the minimax convergence rate of the estimation error is derived under random Gaussian design setting. In simulation studies, we demonstrate followings: (I) under random initializations, solutions of (2) via WMVR-ADMM algorithm converge to a certain estimator whose objective values are same (II) the WNN method outperforms SNN (Yuan et al., 2007) and ANN (Chen et al., 2013) under synthetic settings, (III) the effect of our suggested weight updating scheme is verified through the comparison with the weight setting by Chen et al. (2013). Lastly the application to the real data set shows the effectiveness of our method. Nonetheless, there are several remaining open questions which require further investigations in the future. We summarize them as follows.

1. A question on whether the non-convex ADMM can achieve the global minimizer of (2) is a huge open question. Although empirical results on the convergence of WMVR-ADMM are provided in section 5, they still cannot verify the converged solution is a global minimizer of (2). We leave both empirical and theoretical justifications on this issue as important open problems. Under SNN setting, it is proved that there exists a primal-dual pair of (2) which satisfies the strong duality (Shang and Kong, 2021). Therefore, the existence of saddle point on $\mathcal{L}_0$ can be ensured, so that the global minimizer of (2) can be proved through the classical techniques in Boyd et al. (2011). Nonetheless, we need further investigation whether these conditions can be used under our WNN setting with non-decreasing weights.
2. As previously mentioned in the remark of Theorem 7, $\mathcal{W} := w_p(2w_p - w_1 + \frac{1}{2})/(w_1 - \frac{1}{2})$ is a sole factor that accounts for the effects of weights in the convergence rate of (19). This result naturally leads us to ask the question; "Under

which pair of (ω_1, ω_p), does the estimator from WNN have a faster convergence rate than the one from SNN?". Under the same choices of tuning parameter $\boldsymbol{\lambda}_n$, a naive way for the comparison is to plug $\omega_1 = \omega_p = 1$ in $\mathcal{W}$. That is, we want to find a pair of (ω_1, ω_p) for which $\mathcal{W} \leq 3$ and $\frac{1}{2} < \omega_1 \leq \cdots \leq \omega_p$. The intersected region is illustrated in Figure 8. From our empirical experiences, the region of (ω_1, ω_p), for which WNN is superior than SNN in terms of estimation, is much larger than it is presented in Figure 8. This problem arises from the tightness of the subset $\mathcal{C}$ we derive in Lemma 6. In order to avoid this problem, we suspect that the different approach from using RSC condition of cost function is needed. A paper Law et al. (2021), recently appeared on arXiv, introduces a technique which takes the advantage of controlling the covering number of projection operators corresponding to the subspaces spanned by the design. They consider a problem of solving nuclear norm penalized least square problem, and their technique is independent from RSC condition. It would be an interesting open problem if their technique can be employed in our problem for obtaining a bigger intersected region than that in Figure 8.

references

Attouch, H., Bolte, J. and Svaiter, B. F. (2013) Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized gauss–seidel methods. *Mathematical Programming*, **137**, 91–129.

Bach, F. R. (2008) Consistency of trace norm minimization. *The Journal of Machine Learning Research*, **9**, 1019–1048.

Boyd, S., Parikh, N. and Chu, E. (2011) *Distributed optimization and statistical learning via the alternating direction method of multipliers*. Now Publishers Inc.

Brereton, R. G. (2003) *Chemometrics: data analysis for the laboratory and chemical plant*. John Wiley & Sons.

Candes, E. J., Wakin, M. B. and Boyd, S. P. (2008) Enhancing sparsity by reweighted ℓ_1 minimization. *Journal of Fourier analysis and applications*, **14**, 877–905.

Chen, K., Dong, H. and Chan, K.-S. (2013) Reduced rank regression via adaptive nuclear norm penalization. *Biometrika*, **100**, 901–920.

Fan, J., Gong, W. and Zhu, Z. (2019) Generalized high-dimensional trace regression via nuclear norm regularization. *Journal of econometrics*, **212**, 177–202.

Fan, J., Wang, W. and Zhu, Z. (2021) A shrinkage principle for heavy-tailed data: High-dimensional robust low-rank matrix recovery. *Annals of statistics*, **49**, 1239.

Golub, G. H., Heath, M. and Wahba, G. (1979) Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, **21**, 215–223.

Gu, S., Xie, Q., Meng, D., Zuo, W., Feng, X. and Zhang, L. (2017) Weighted nuclear norm minimization and its applications to low level vision. *International journal of computer vision*, **121**, 183–208.

Gu, S., Zhang, L., Zuo, W. and Feng, X. (2014) Weighted nuclear norm minimization with application to image denoising. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2862–2869.

Isenmann, A. (2008) Modern multivariate statistical techniques. *Regression, Classification and Manifold Learning*, **25**, 733.

Kim, G., Cho, J. and Kang, M. (2020) Cauchy noise removal by weighted nuclear norm minimization. *Journal of Scientific Computing*, **83**, 1–21.

Koltchinskii, V., Lounici, K. and Tsybakov, A. B. (2011) Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, **39**, 2302–2329.

Law, M., Ritov, Y., Zhang, R. and Zhu, Z. (2021) Rank-constrained least-squares: Prediction and inference. *arXiv preprint arXiv:2111.14287*.

Lee, J. D., Sun, Y. and Taylor, J. E. (2015) On model selection consistency of regularized m-estimators. *Electronic Journal of Statistics*, **9**, 608–642.

Liu, S., Hu, Q., Li, P., Zhao, J., Liu, M. and Zhu, Z. (2018) Speckle suppression based on weighted nuclear norm minimization and grey theory. *IEEE Transactions on Geoscience and Remote Sensing*, **57**, 2700–2708.

Mirsky, L. (1975) A trace inequality of john von neumann. *Monatshefte für mathematik*, **79**, 303–306.

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.

Rohde, A. and Tsybakov, A. B. (2011) Estimation of high-dimensional low-rank matrices. *The Annals of Statistics*, **39**, 887–930.

Ruhe, A. (1970) Perturbation bounds for means of eigenvalues and invariant subspaces. *BIT Numerical Mathematics*, **10**, 343–354.

Shang, P. and Kong, L. (2021) Regularization parameter selection for the low rank matrix recovery. *Journal of Optimization Theory and Applications*, **189**, 772–792.

Sun, T., Jiang, H. and Cheng, L. (2017) Global convergence of proximal iteratively reweighted algorithm. *Journal of Global Optimization*, **68**, 815–826.

Von Neumann, J. (1937) *Some matrix-inequalities and metrization of matric space*. JSTOR.

Wainwright, M. J. (2019) *High-dimensional statistics: A non-asymptotic viewpoint*, vol. 48. Cambridge University Press.

Wang, Y., Yin, W. and Zeng, J. (2019) Global convergence of admm in nonconvex nonsmooth optimization. *Journal of Scientific Computing*, **78**, 29–63.

Xu, J., Zhang, L., Zhang, D. and Feng, X. (2017) Multi-channel weighted nuclear norm minimization for real color image denoising. In *Proceedings of the IEEE international conference on computer vision*, 1096–1104.

Yair, N. and Michaeli, T. (2018) Multi-scale weighted nuclear norm image restoration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3165–3174.

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.

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

8 | APPENDIX

In this section, an extended algorithm of WMVR-ADMM to trace regression model is proposed in Section 8.1. Subsequently, proofs on Lemma 3, Lemma 4, Theorem 2, Lemma 6, and Theorem 7 are provided in Sections 8.2, 8.3, 8.4, and 8.5, respectively.

8.1 | Extension of WMVR-ADMM to Trace Regression Model

First, let us consider a following trace regression problem :

$$y_i = \mathbf{tr}(\boldsymbol{X}_i^T \boldsymbol{\Theta}^\star) + \varepsilon_i \ , \ i = 1, \cdots, n,$$

where $\boldsymbol{X}_i \in \boldsymbol{R}^{d_1 \times d_2}$ is a known measurement matrix for $i = 1, \cdots, n$ and $\{\varepsilon_i\}_{i=1}^n \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, \sigma^2)$. In this subsection, we present an extension of WMVR-ADMM algorithm for solving a following optimization problem.

$$\min_{\boldsymbol{\Theta}} \left\{ \frac{1}{2n} \sum_{i=1}^{n} (y_i - \mathbf{tr}(\boldsymbol{X}_i^T \boldsymbol{\Theta}))^2 + \boldsymbol{\lambda}_n ||\boldsymbol{\Theta}||_{\omega,\star} \right\}.$$

First, let $v(M) \in \mathbb{R}^{d_1 d_2}$ be the vectorized version of matrix M concatenating columns of $M \in \mathbb{R}^{d_1 \times d_2}$ into one column vector, and let us also define an inverse operator $\mathbf{Mat}\big[v(M)\big] := M$. With this notation, the algorithm is summarized in the Algorithm 2. The presented algorithm can be derived easily by using exactly the same techniques employed in subsection 2.1 by plugging $f(\boldsymbol{\Theta}) := -\frac{1}{n}\mathbf{tr}\big(\big(\sum_{j=1}^n y_i \boldsymbol{X}_i\big)^\top \boldsymbol{\Theta}\big) + \boldsymbol{\lambda}_n ||\boldsymbol{\Theta}||_{\omega,\star}$ and $g(\boldsymbol{\Gamma}) = \frac{1}{2n}\sum_{j=1}^n \mathbf{tr}\big(\boldsymbol{X}_{\boldsymbol{i}}^\top \boldsymbol{\Gamma}\big)^2$ in (4), and we omit the detailed derivation.

```
Algorithm 2: ADMM for weighted Trace Regression. (WTR-ADMM)
Input : $\{\boldsymbol{X}_i, y_i\}_{i=1}^n$, $\boldsymbol{\lambda}_n \geq 0$.
Prelimiaries : $\boldsymbol{M_{y x}} := \sum_{i=1}^n y_i \boldsymbol{X}_i$, and $\mathcal{A} := \frac{1}{n}\sum_{i=1}^n v(\boldsymbol{X}_i)v(\boldsymbol{X}_i)^\top + \rho \cdot \mathcal{I}_{d_1 d_2 \times d_1 d_2}$.
Initialization : $\boldsymbol{\Theta}^{(0)} = \mathbf{0}$, $\boldsymbol{\Gamma}^{(0)} = \mathbf{0}$, $\boldsymbol{\Lambda}^{(0)} = \mathbf{0} \in \mathbb{R}^{d_1 \times d_2}$.
    Repeat following Steps :
        Step 1. Let $\boldsymbol{B^{(k)}} := \frac{1}{n}\boldsymbol{M_{y x}} - \boldsymbol{\Lambda}^{(k)} + \rho \cdot \boldsymbol{\Gamma}^{(k)}$.    $\boldsymbol{B^{(k)}} = \boldsymbol{U}^{\mathbf{B}}\boldsymbol{D}^{\mathbf{B}}(\boldsymbol{V}^{\mathbf{B}})^\top$. (SVD)
                Set $\mathcal{S}_{\boldsymbol{\lambda}_n \omega}(\boldsymbol{D}^{\mathbf{B}}) = \mathbf{diag}\Big\{ \max\Big\{\frac{1}{\rho}(\sigma_j(\boldsymbol{B}^{(k)}) - \boldsymbol{\lambda}_n w_j), 0\Big\} \text{ for } j = 1, \ldots, p\Big\}$.
                $\boldsymbol{\Theta}^{(k+1)} = \boldsymbol{U}^{\mathbf{B}}\mathcal{S}_{\boldsymbol{\lambda}_n \omega}(\boldsymbol{D}^{\mathbf{B}})(\boldsymbol{V}^{\mathbf{B}})^\top$.
        Step 2. $\boldsymbol{\Gamma}^{(k+1)} = \mathbf{Mat}\big[\mathcal{A}^{-1}(\rho v(\boldsymbol{\Theta}^{(k+1)}) - v(\boldsymbol{\Lambda}^{(k)}))\big]$.
        Step 3. $\boldsymbol{\Lambda}^{(k+1)} = \boldsymbol{\Lambda}^{(k)} + \rho(\boldsymbol{\Theta}^{(k+1)} - \boldsymbol{\Gamma}^{(k+1)})$.
    Until $||\boldsymbol{\Theta}^{(k+1)} - \boldsymbol{\Gamma}^{(k+1)}||_F \leq 10^{-7}$ and $||\boldsymbol{\Gamma}^{(k+1)} - \boldsymbol{\Gamma}^{(k)}||_F \leq 10^{-7}$.
Output : $\widehat{\boldsymbol{\Theta}} = \boldsymbol{\Theta}^{(k+1)}$.
```

8.2 | Proof of Lemma 3

By the result of Lemma 1, we have

$$\mathcal{L}_\rho\big(\boldsymbol{\Theta}^{(k)}, \boldsymbol{\Gamma}^{(k)}, \boldsymbol{\Lambda}^{(k)}\big) - \mathcal{L}_\rho\big(\boldsymbol{\Theta}^{(k+1)}, \boldsymbol{\Gamma}^{(k)}, \boldsymbol{\Lambda}^{(k)}\big) \geq 0. \tag{31}$$

Now, we control the following difference term.

$$\begin{aligned}
&\mathcal{L}_\rho(\boldsymbol{\Theta}^{(k+1)}, \boldsymbol{\Gamma}^{(k)}, \boldsymbol{\Lambda}^{(k)}) - \mathcal{L}_\rho(\boldsymbol{\Theta}^{(k+1)}, \boldsymbol{\Gamma}^{(k+1)}, \boldsymbol{\Lambda}^{(k)}) \\
&\quad = g(\boldsymbol{\Gamma}^{(k)}) - g(\boldsymbol{\Gamma}^{(k+1)}) - \mathbf{tr}(\boldsymbol{\Lambda}^{(k)\top}(\boldsymbol{\Gamma}^{(k)} - \boldsymbol{\Gamma}^{(k+1)})) \\
&\qquad\quad - \rho \cdot \mathbf{tr}((\boldsymbol{\Theta}^{(k+1)} - \boldsymbol{\Gamma}^{(k+1)})^\top(\boldsymbol{\Gamma}^{(k)} - \boldsymbol{\Gamma}^{(k+1)})) + \frac{\rho}{2}\left\|\boldsymbol{\Gamma}^{(k)} - \boldsymbol{\Gamma}^{(k+1)}\right\|_F^2 \\
&\quad = g(\boldsymbol{\Gamma}^{(k)}) - g(\boldsymbol{\Gamma}^{(k+1)}) - \mathbf{tr}(\boldsymbol{\Lambda}^{(k+1)\top}(\boldsymbol{\Gamma}^{(k)} - \boldsymbol{\Gamma}^{(k+1)})) + \frac{\rho}{2}\left\|\boldsymbol{\Gamma}^{(k)} - \boldsymbol{\Gamma}^{(k+1)}\right\|_F^2. \qquad (32)
\end{aligned}$$

Note that we use $\boldsymbol{\Lambda}^{(k+1)} = \boldsymbol{\Lambda}^{(k)} + \rho(\boldsymbol{\Theta}^{(k+1)} - \boldsymbol{\Gamma}^{(k+1)})$ in the last equality. Recall the definition of $\mathcal{L}_\rho(\boldsymbol{\Theta}^{(k)}, \boldsymbol{\Gamma}^{(k)}, \boldsymbol{\Lambda}^{(k)})$ from (5) and $\boldsymbol{\Lambda}^{(k+1)} = \boldsymbol{\Lambda}^{(k)} + \rho(\boldsymbol{\Theta}^{(k+1)} - \boldsymbol{\Gamma}^{(k+1)})$. Then, we have

$$\mathcal{L}_\rho(\boldsymbol{\Theta}^{(k+1)}, \boldsymbol{\Gamma}^{(k+1)}, \boldsymbol{\Lambda}^{(k)}) - \mathcal{L}_\rho(\boldsymbol{\Theta}^{(k+1)}, \boldsymbol{\Gamma}^{(k+1)}, \boldsymbol{\Lambda}^{(k+1)}) = -\frac{1}{\rho}\left\|\boldsymbol{\Lambda}^{(k)} - \boldsymbol{\Lambda}^{(k+1)}\right\|_F^2. \qquad (33)$$

By combining (32) and (33), we have

$$\begin{aligned}
&\mathcal{L}_\rho(\boldsymbol{\Theta}^{(k+1)}, \boldsymbol{\Gamma}^{(k)}, \boldsymbol{\Lambda}^{(k)}) - \mathcal{L}_\rho(\boldsymbol{\Theta}^{(k+1)}, \boldsymbol{\Gamma}^{(k+1)}, \boldsymbol{\Lambda}^{(k+1)}) \\
&\quad = g(\boldsymbol{\Gamma}^{(k)}) - g(\boldsymbol{\Gamma}^{(k+1)}) - \mathbf{tr}(\boldsymbol{\Lambda}^{(k+1)\top}(\boldsymbol{\Gamma}^{(k)} - \boldsymbol{\Gamma}^{(k+1)})) \\
&\qquad\qquad + \frac{\rho}{2}\left\|\boldsymbol{\Gamma}^{(k)} - \boldsymbol{\Gamma}^{(k+1)}\right\|_F^2 - \frac{1}{\rho}\left\|\boldsymbol{\Lambda}^{(k)} - \boldsymbol{\Lambda}^{(k+1)}\right\|_F^2. \qquad (34)
\end{aligned}$$

Recall the definition of $\boldsymbol{\Gamma}^{(k+1)}$ from **Step 2** of WMVR-ADMM.

$$\boldsymbol{\Gamma}^{(k+1)} = \underset{\boldsymbol{\Gamma} \in \mathbb{R}^{d_1 \times d_2}}{\operatorname{argmin}} \left\{ g(\boldsymbol{\Gamma}) - \mathbf{tr}(\boldsymbol{\Lambda}^{(k)\top}\boldsymbol{\Gamma}) + \frac{\rho}{2}\left\|\boldsymbol{\Gamma} - \boldsymbol{\Theta}^{(k+1)}\right\|_F^2 \right\}.$$

Since $\boldsymbol{\Gamma}^{(k+1)}$ is a stationary point of the above optimization problem, we have

$$\nabla g(\boldsymbol{\Gamma}^{(k+1)}) = \boldsymbol{\Lambda}^{(k)} + \rho(\boldsymbol{\Theta}^{(k+1)} - \boldsymbol{\Gamma}^{(k+1)}) = \boldsymbol{\Lambda}^{(k+1)},$$

where $\nabla g(\cdot)$ is a gradient of g. Likewise, we get $\nabla g(\boldsymbol{\Gamma}^{(k)}) = \boldsymbol{\Lambda}^{(k)}$. Recall the definition of $g(\cdot)$, then we can easily have

$$\left\|\boldsymbol{\Lambda}^{(k+1)} - \boldsymbol{\Lambda}^{(k)}\right\|_F = \left\|\nabla g(\boldsymbol{\Gamma}^{(k+1)}) - \nabla g(\boldsymbol{\Gamma}^{(k)})\right\|_F \leq \sigma_1\left(\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}\right) \cdot \left\|\boldsymbol{\Gamma}^{(k+1)} - \boldsymbol{\Gamma}^{(k)}\right\|_F. \qquad (35)$$

Function g is Lipschitz smooth with constant $L_{\nabla g} := \sigma_1\left(\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}\right)$. Then, we have

$$g(\boldsymbol{\Gamma}^{(k)}) - g(\boldsymbol{\Gamma}^{(k+1)}) - \mathbf{tr}\Big(\nabla g(\boldsymbol{\Gamma}^{(k+1)})^\top(\boldsymbol{\Gamma}^{(k)} - \boldsymbol{\Gamma}^{(k+1)})\Big) \geq -\frac{L_{\nabla g}}{2}\left\|\boldsymbol{\Gamma}^{(k+1)} - \boldsymbol{\Gamma}^{(k)}\right\|_F^2. \qquad (36)$$

Recall $\nabla g(\boldsymbol{\Gamma}^{(k+1)}) = \boldsymbol{\Lambda}^{(k+1)}$, combining (31), (34), (35), and (36) yields

$$\mathcal{L}_\rho(\boldsymbol{\Theta}^{(k)}, \boldsymbol{\Gamma}^{(k)}, \boldsymbol{\Lambda}^{(k)}) - \mathcal{L}_\rho(\boldsymbol{\Theta}^{(k+1)}, \boldsymbol{\Gamma}^{(k+1)}, \boldsymbol{\Lambda}^{(k+1)}) \geq \Big(-\frac{L_{\nabla g}}{2} - \frac{1}{\rho}L_{\nabla g}^2 + \frac{\rho}{2}\Big) \cdot \left\|\boldsymbol{\Gamma}^{(k+1)} - \boldsymbol{\Gamma}^{(k)}\right\|_F^2.$$

Setting $\rho > 2L_{\nabla g}$ makes $C_1 := -\frac{L_{\nabla g}}{2} - \frac{1}{\rho}L_{\nabla g}^2 + \frac{\rho}{2} > 0$, which implies that $\mathcal{L}_\rho(\boldsymbol{\Theta}^{(k)}, \boldsymbol{\Gamma}^{(k)}, \boldsymbol{\Lambda}^{(k)})$ is non-increasing over $k \in \mathbb{R} \cup \{0\}$. Now, we will prove $\mathcal{L}_\rho(\boldsymbol{\Theta}^{(k)}, \boldsymbol{\Gamma}^{(k)}, \boldsymbol{\Lambda}^{(k)})$ is bounded below over $k \in \mathbb{N} \cup \{0\}$.

$$\begin{aligned}
\mathcal{L}_\rho(\boldsymbol{\Theta}^{(k)}, \boldsymbol{\Gamma}^{(k)}, \boldsymbol{\Lambda}^{(k)}) &= f(\boldsymbol{\Theta}^{(k)}) + g(\boldsymbol{\Gamma}^{(k)}) + \mathbf{tr}(\boldsymbol{\Lambda}^{(k)\top}(\boldsymbol{\Theta}^{(k)} - \boldsymbol{\Gamma}^{(k)})) + \frac{\rho}{2}||\boldsymbol{\Theta}^{(k)} - \boldsymbol{\Gamma}^{(k)}||_{\mathsf{F}}^2 \\
&= f(\boldsymbol{\Theta}^{(k)}) + g(\boldsymbol{\Gamma}^{(k)}) + \mathbf{tr}(\nabla g(\boldsymbol{\Gamma}^{(k)})^\top(\boldsymbol{\Theta}^{(k)} - \boldsymbol{\Gamma}^{(k)})) + \frac{\rho}{2}||\boldsymbol{\Theta}^{(k)} - \boldsymbol{\Gamma}^{(k)}||_{\mathsf{F}}^2 \\
&\geq g(\boldsymbol{\Gamma}^{(k)}) + \mathbf{tr}(\nabla g(\boldsymbol{\Gamma}^{(k)})^\top(\boldsymbol{\Theta}^{(k)} - \boldsymbol{\Gamma}^{(k)})) + \frac{\rho}{2}||\boldsymbol{\Theta}^{(k)} - \boldsymbol{\Gamma}^{(k)}||_{\mathsf{F}}^2 \\
&\geq g(\boldsymbol{\Theta}^{(k)}) - \frac{L_{\nabla g}}{2}||\boldsymbol{\Theta}^{(k)} - \boldsymbol{\Gamma}^{(k)}||_{\mathsf{F}}^2 + \frac{\rho}{2}||\boldsymbol{\Theta}^{(k)} - \boldsymbol{\Gamma}^{(k)}||_{\mathsf{F}}^2 \\
&\geq g(\boldsymbol{\Theta}^{(k)}) := \frac{1}{2n}\left\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\Theta}^{(k)}\right\|_{\mathsf{F}}^2.
\end{aligned}$$

In the first inequality, $f(\boldsymbol{\Theta}^{(k)}) \geq 0$ is used. In the second inequality, Lipschitz smoothness of g with constant $L_{\nabla g}$ is used, and in the last inequality, the choice on $\rho > 2L_{\nabla g}$ is used. It is obvious that $g(\boldsymbol{\Theta}^{(k)})$ is bounded below from 0.

As long as $\{(\boldsymbol{\Theta}^{(0)}, \boldsymbol{\Gamma}^{(0)}, \boldsymbol{\Lambda}^{(0)})\}$ is bounded, it is easy to see the generated sequence $\{\boldsymbol{\Theta}^{(k)}, \boldsymbol{\Gamma}^{(k)}, \boldsymbol{\Lambda}^{(k)}\}_{k\geq 1}$ is bounded as well. Since the minimizers of **Step 1.** and **Step 2.** have explicit closed form solution, the pair $\{\boldsymbol{\Theta}^{(1)}, \boldsymbol{\Gamma}^{(1)}\}$ is bounded, and by **Step 3.**, the boundedness of $\boldsymbol{\Lambda}^{(1)}$ is automatically ensured. Applying the same logic over the $k \geq 2$ yields the claim. □

8.3 | Proof of Lemma 4

Let us define the partial derivative of $\mathcal{L}_\rho(\boldsymbol{\Theta}^{(k+1)}, \boldsymbol{\Gamma}^{(k+1)}, \boldsymbol{\Lambda}^{(k+1)})$ as

$$\partial\mathcal{L}_\rho(\boldsymbol{\Theta}^{(k+1)}, \boldsymbol{\Gamma}^{(k+1)}, \boldsymbol{\Lambda}^{(k+1)}) := (\partial_\Theta\mathcal{L}_\rho, \nabla_\Gamma\mathcal{L}_\rho, \nabla_\Lambda\mathcal{L}_\rho)(\boldsymbol{\Theta}^{(k+1)}, \boldsymbol{\Gamma}^{(k+1)}, \boldsymbol{\Lambda}^{(k+1)}).$$

It is easy to see followings:

$$\begin{aligned}
\nabla_\Gamma\mathcal{L}_\rho(\boldsymbol{\Theta}^{(k+1)}, \boldsymbol{\Gamma}^{(k+1)}, \boldsymbol{\Lambda}^{(k+1)}) &= \boldsymbol{\Lambda}^{(k+1)} - \boldsymbol{\Lambda}^{(k)} \\
\nabla_\Lambda\mathcal{L}_\rho(\boldsymbol{\Theta}^{(k+1)}, \boldsymbol{\Gamma}^{(k+1)}, \boldsymbol{\Lambda}^{(k+1)}) &= \frac{1}{\rho}(\boldsymbol{\Lambda}^{(k+1)} - \boldsymbol{\Lambda}^{(k)}).
\end{aligned}$$

Since $\boldsymbol{\Theta}^{(k+1)}$ is a minimizer of **Step 1.**, it satisfies a following stationary condition.

$$\mathbf{0} \in \partial f(\boldsymbol{\Theta}^{(k+1)}) + \boldsymbol{\Lambda}^{(k)} + \rho(\boldsymbol{\Theta}^{(k+1)} - \boldsymbol{\Gamma}^{(k)}). \tag{37}$$

Then, we are interested in getting a subdifferential of $\mathcal{L}_\rho(\boldsymbol{\Theta}^{(k+1)}, \boldsymbol{\Gamma}^{(k+1)}, \boldsymbol{\Lambda}^{(k+1)})$ with respect to $\boldsymbol{\Theta}$, which can be calculated as follows:

$$\begin{aligned}
\partial_{\boldsymbol{\Theta}}\mathcal{L}_\rho(\boldsymbol{\Theta}^{(k+1)}, \boldsymbol{\Gamma}^{(k+1)}, \boldsymbol{\Lambda}^{(k+1)}) &= \partial f(\boldsymbol{\Theta}^{(k+1)}) + \boldsymbol{\Lambda}^{(k+1)} + \rho(\boldsymbol{\Theta}^{(k+1)} - \boldsymbol{\Gamma}^{(k+1)}) \\
&= \partial f(\boldsymbol{\Theta}^{(k+1)}) + \boldsymbol{\Lambda}^{(k)} + \rho(\boldsymbol{\Theta}^{(k+1)} - \boldsymbol{\Gamma}^{(k)}) + (\boldsymbol{\Lambda}^{(k+1)} - \boldsymbol{\Lambda}^{(k)}) + \rho(\boldsymbol{\Gamma}^{(k)} - \boldsymbol{\Gamma}^{(k+1)}).
\end{aligned}$$

Then, by (37), we have

$$(\boldsymbol{\Lambda}^{(k+1)} - \boldsymbol{\Lambda}^{(k)}) + \rho(\boldsymbol{\Gamma}^{(k)} - \boldsymbol{\Gamma}^{(k+1)}) \in \partial_{\boldsymbol{\Theta}}\mathcal{L}_\rho(\boldsymbol{\Theta}^{(k+1)}, \boldsymbol{\Gamma}^{(k+1)}, \boldsymbol{\Lambda}^{(k+1)}).$$

If we define $p^{(k+1)}$ as

$$p^{(k+1)} := \Big((\boldsymbol{\Lambda}^{(k+1)} - \boldsymbol{\Lambda}^{(k)}) + \rho(\boldsymbol{\Gamma}^{(k)} - \boldsymbol{\Gamma}^{(k+1)}), \boldsymbol{\Lambda}^{(k+1)} - \boldsymbol{\Lambda}^{(k)}, \frac{1}{\rho}(\boldsymbol{\Lambda}^{(k+1)} - \boldsymbol{\Lambda}^{(k)})\Big), \tag{38}$$

then $p^{(k+1)} \in \partial\mathcal{L}_\rho(\boldsymbol{\Theta}^{(k+1)}, \boldsymbol{\Gamma}^{(k+1)}, \boldsymbol{\Lambda}^{(k+1)})$. Furthermore, its Frobenious norm can be bounded by combining (35) and (38) as follows:

$$\left\|p^{(k+1)}\right\|_{\mathrm{F}} \le \Big(\rho + (2 + \frac{1}{\rho})L_{\nabla g}\Big) \cdot \left\|\boldsymbol{\Gamma}^{(k+1)} - \boldsymbol{\Gamma}^{(k)}\right\|_{\mathrm{F}}.$$

Setting $C_2 := \rho + (2 + \frac{1}{\rho})L_{\nabla g}$ yields the claim. □

8.4 | Proof of Lemma 6

Since $\widehat{\boldsymbol{\Theta}}$ is a minimizer and $\boldsymbol{\Theta}^\star$ is a feasible solution of the optimization problem in (2) respectively, we have a following basic inequality:

$$\frac{1}{2n}\left\|\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\Theta}}\right\|_{\mathrm{F}}^2 + \boldsymbol{\lambda}_n||\widehat{\boldsymbol{\Theta}}||_{w,\star} \le \frac{1}{2n}\left\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\Theta}^\star\right\|_{\mathrm{F}}^2 + \boldsymbol{\lambda}_n||\boldsymbol{\Theta}^\star||_{w,\star}. \tag{39}$$

Plugging in $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\Theta}^\star + \boldsymbol{E}$ in the (39) yields

$$\frac{1}{2n}\left\|\boldsymbol{X}(\boldsymbol{\Theta}^\star - \widehat{\boldsymbol{\Theta}})\right\|_{\mathrm{F}}^2 \le \frac{1}{n}\mathbf{tr}((\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^\star)^\top \boldsymbol{X}^\top \boldsymbol{E}) + \boldsymbol{\lambda}_n(||\boldsymbol{\Theta}^\star||_{w,\star} - ||\widehat{\boldsymbol{\Theta}}||_{w,\star}). \tag{40}$$

By denoting $\widehat{\boldsymbol{\Delta}} \equiv \widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^\star$ and since left-hand side of (40) is ≥ 0, we have

$$0 \le \frac{1}{n}\mathbf{tr}(\widehat{\boldsymbol{\Delta}}^\top \boldsymbol{X}^\top \boldsymbol{E}) + \boldsymbol{\lambda}_n(||\boldsymbol{\Theta}^\star||_{w,\star} - ||\widehat{\boldsymbol{\Delta}} + \boldsymbol{\Theta}^\star||_{w,\star}). \tag{41}$$

First, we will control the upper-bound on the second term of the (41). By the definition of the weighted nuclear norm, we can re-write the term as follows:

$$\begin{aligned}
||\boldsymbol{\Theta}^\star||_{w,\star} - ||\widehat{\boldsymbol{\Delta}} + \boldsymbol{\Theta}^\star||_{w,\star} &= \sum_{j=1}^{p} w_j\sigma_j(\boldsymbol{\Theta}^\star) - \sum_{j=1}^{p} w_j\sigma_j(\widehat{\boldsymbol{\Delta}} + \boldsymbol{\Theta}^\star) \\
&= \left[w_p\sum_{j=1}^{p}\sigma_j(\boldsymbol{\Theta}^\star) - \sum_{j=1}^{p}(w_p - w_j)\sigma_j(\boldsymbol{\Theta}^\star)\right] - \left[w_p\sum_{j=1}^{p}\sigma_j(\widehat{\boldsymbol{\Delta}} + \boldsymbol{\Theta}^\star) - \sum_{j=1}^{p}(w_p - w_j)\sigma_j(\widehat{\boldsymbol{\Delta}} + \boldsymbol{\Theta}^\star)\right] \\
&= w_p\left[\sum_{j=1}^{p}\sigma_j(\boldsymbol{\Theta}^\star) - \sum_{j=1}^{p}\sigma_j(\widehat{\boldsymbol{\Delta}} + \boldsymbol{\Theta}^\star)\right] + \left[\sum_{j=1}^{p}(w_p - w_j)\{\sigma_j(\widehat{\boldsymbol{\Delta}} + \boldsymbol{\Theta}^\star) - \sigma_j(\boldsymbol{\Theta}^\star)\}\right] \\
&= \underbrace{w_p(||\boldsymbol{\Theta}^\star||_\star - ||\widehat{\boldsymbol{\Delta}} + \boldsymbol{\Theta}^\star||_\star)}_{:=\mathrm{I}} + \underbrace{(||\widehat{\boldsymbol{\Delta}} + \boldsymbol{\Theta}^\star||_{w_p - w,\star} - ||\boldsymbol{\Theta}^\star||_{w_p - w,\star})}_{:=\mathrm{II}},
\end{aligned} \tag{42}$$

where $||\Theta||_{w_p-w,\star} = \sum_{j=1}^{p}(w_p - w_j)\sigma_j(\boldsymbol{\Theta})$.

Recall the definitions of the two subspaces $\mathcal{M}_r$ and $\overline{\mathcal{M}_r}^{\perp}$ in subsection 3.2.2. For any $r \le p$, we have

$$\boldsymbol{\Theta}^{\star} = \Pi_{\mathcal{M}_r}(\boldsymbol{\Theta}^{\star}) + \Pi_{\overline{\mathcal{M}_r}^{\perp}}(\boldsymbol{\Theta}^{\star}). \tag{43}$$

Recall that $\widehat{\boldsymbol{\Delta}}'' \in \Pi_{\overline{\mathcal{M}_r}^{\perp}}(\widehat{\boldsymbol{\Delta}})$ and $\widehat{\boldsymbol{\Delta}}' = \widehat{\boldsymbol{\Delta}} - \widehat{\boldsymbol{\Delta}}''$. Then, we can control the term $||\widehat{\boldsymbol{\Delta}} + \boldsymbol{\Theta}^{\star}||_{\star}$ as follows:

$$\begin{aligned}
||\widehat{\boldsymbol{\Delta}} + \boldsymbol{\Theta}^{\star}||_{\star} &= ||\widehat{\boldsymbol{\Delta}}' + \widehat{\boldsymbol{\Delta}}'' + \Pi_{\mathcal{M}_r}(\boldsymbol{\Theta}^{\star}) + \Pi_{\overline{\mathcal{M}_r}^{\perp}}(\boldsymbol{\Theta}^{\star})||_{\star} \\
&\ge ||\widehat{\boldsymbol{\Delta}}'' + \Pi_{\mathcal{M}_r}(\boldsymbol{\Theta}^{\star})||_{\star} - \{||\widehat{\boldsymbol{\Delta}}'||_{\star} + ||\Pi_{\overline{\mathcal{M}_r}^{\perp}}(\boldsymbol{\Theta}^{\star})||_{\star}\} \\
&= ||\widehat{\boldsymbol{\Delta}}''||_{\star} + ||\Pi_{\mathcal{M}_r}(\boldsymbol{\Theta}^{\star})||_{\star} - \{||\widehat{\boldsymbol{\Delta}}'||_{\star} + ||\Pi_{\overline{\mathcal{M}_r}^{\perp}}(\boldsymbol{\Theta}^{\star})||_{\star}\},
\end{aligned} \tag{44}$$

where in the first inequality, we used the triangle inequality of $||\cdot||_{\star}$ and in the last equality, the decomposability of $||\cdot||_{\star}$ with respect to a pair of subspaces $(\mathcal{M}_r, \overline{\mathcal{M}_r}^{\perp})$ is used. With (44), we are ready to control the term I in (42).

$$\begin{aligned}
&w_p\Big(||\boldsymbol{\Theta}^{\star}||_{\star} - ||\widehat{\boldsymbol{\Delta}} + \boldsymbol{\Theta}^{\star}||_{\star}\Big) \\
&\quad\le w_p \cdot \Bigg\{ \Big(||\Pi_{\mathcal{M}_r}(\boldsymbol{\Theta}^{\star})||_{\star} + ||\Pi_{\overline{\mathcal{M}_r}^{\perp}}(\boldsymbol{\Theta}^{\star})||_{\star}\Big) \\
&\qquad\qquad - \Big(||\widehat{\boldsymbol{\Delta}}''||_{\star} + ||\Pi_{\mathcal{M}_r}(\boldsymbol{\Theta}^{\star})||_{\star} - \{||\widehat{\boldsymbol{\Delta}}'||_{\star} + ||\Pi_{\overline{\mathcal{M}_r}^{\perp}}(\boldsymbol{\Theta}^{\star})||_{\star}\}\Big) \Bigg\} \\
&\quad= w_p \cdot \Big\{ 2||\Pi_{\overline{\mathcal{M}_r}^{\perp}}(\boldsymbol{\Theta}^{\star})||_{\star} + ||\widehat{\boldsymbol{\Delta}}'||_{\star} - ||\widehat{\boldsymbol{\Delta}}''||_{\star} \Big\}
\end{aligned} \tag{45}$$

Note that the equality $||\Theta^{\star}||_{\star} = ||\Pi_{\mathcal{M}_r}(\boldsymbol{\Theta}^{\star})||_{\star} + ||\Pi_{\overline{\mathcal{M}_r}^{\perp}}(\boldsymbol{\Theta}^{\star})||_{\star}$ is used in the first inequality due to (43).

Now the term II in (42) needs to be controlled. First, we need to see the norm $||\cdot||_{w_p-w,\star} = \sum_{j=1}^{p}(w_p - w_j)\sigma_j(\cdot)$ with respect to any pair of matrices: $(A, B) \in (\mathcal{M}_r, \overline{\mathcal{M}_r}^{\perp})$ satisfies the decomposability, meaning $||A + B||_{w_p-w,\star} = ||A||_{w_p-w,\star} + ||B||_{w_p-w,\star}$. By definition of the subspace pair $(\mathcal{M}_r, \overline{\mathcal{M}_r}^{\perp})$, we can write A and B as

$$A = \boldsymbol{U}\begin{bmatrix} \boldsymbol{T}_{1,1} & \mathbf{0}_{r\times(p-r)} \\ \mathbf{0}_{(p-r)\times r} & \mathbf{0}_{(p-r)\times(p-r)} \end{bmatrix}\boldsymbol{V}^{\top}, \qquad B = \boldsymbol{U}\begin{bmatrix} \mathbf{0}_{r\times r} & \mathbf{0}_{r\times(p-r)} \\ \mathbf{0}_{(p-r)\times r} & \boldsymbol{T}_{2,2} \end{bmatrix}\boldsymbol{V}^{\top},$$

where $\boldsymbol{T}_{1,1} \in \mathbb{R}^{r\times r}$ and $\boldsymbol{T}_{2,2} \in \mathbb{R}^{(p-r)\times(p-r)}$ are arbitrary matrices. Define two diagonal matrices $\boldsymbol{W}_1 := \text{diag}(w_p - w_1, \ldots, w_p - w_r)$ and $\boldsymbol{W}_2 := \text{diag}(w_p - w_{r+1}, \ldots, w_p - w_p)$. Then, we have

$$\begin{aligned}
||A + B||_{w_p-w,\star} &= \left\|\begin{bmatrix} \boldsymbol{W}_1\boldsymbol{T}_{1,1} & \mathbf{0}_{r\times(p-r)} \\ \mathbf{0}_{(p-r)\times r} & \mathbf{0}_{(p-r)\times(p-r)} \end{bmatrix} + \begin{bmatrix} \mathbf{0}_{r\times r} & \mathbf{0}_{r\times(p-r)} \\ \mathbf{0}_{(p-r)\times r} & \boldsymbol{W}_2\boldsymbol{T}_{2,2} \end{bmatrix}\right\|_{\star} \\
&= \left\|\begin{bmatrix} \boldsymbol{W}_1\boldsymbol{T}_{1,1} & \mathbf{0}_{r\times(p-r)} \\ \mathbf{0}_{(p-r)\times r} & \mathbf{0}_{(p-r)\times(p-r)} \end{bmatrix}\right\|_{\star} + \left\|\begin{bmatrix} \mathbf{0}_{r\times r} & \mathbf{0}_{r\times(p-r)} \\ \mathbf{0}_{(p-r)\times r} & \boldsymbol{W}_2\boldsymbol{T}_{2,2} \end{bmatrix}\right\|_{\star} \\
&= ||A||_{w_p-w,\star} + ||B||_{w_p-w,\star}.
\end{aligned}$$

In the first equality, the definition of $\|\cdot\|_{w_p-w,\star}$ and the invariance of the nuclear norm to orthogonal transformation to multiplication by the matrices $\boldsymbol{U}^{\star}$ and $\boldsymbol{V}^{\star}$ are used.

Using this fact, similarly with (44) and (45), we get the upper-bound on II in the equality (42):

$$\begin{aligned}
||\widehat{\boldsymbol{\Delta}}+\boldsymbol{\Theta}^{\star}||_{w_p-w,\star}-||\boldsymbol{\Theta}^{\star}||_{w_p-w,\star} &\Leftrightarrow ||\Pi_{\mathcal{M}_r}(\boldsymbol{\Theta}^{\star})+\Pi_{\overline{\mathcal{M}}_r^{\perp}}(\boldsymbol{\Theta}^{\star})+\widehat{\boldsymbol{\Delta}}'+\widehat{\boldsymbol{\Delta}}''||_{w_p-w,\star}-||\boldsymbol{\Theta}^{\star}||_{w_p-w,\star} \\
&\leq ||\Pi_{\mathcal{M}_r}(\boldsymbol{\Theta}^{\star})+\widehat{\boldsymbol{\Delta}}''||_{w_p-w,\star}+||\Pi_{\overline{\mathcal{M}}_r^{\perp}}(\boldsymbol{\Theta}^{\star})||_{w_p-w,\star}+||\widehat{\boldsymbol{\Delta}}'||_{w_p-w,\star}-||\boldsymbol{\Theta}^{\star}||_{w_p-w,\star} \\
&= \Big\{||\Pi_{\mathcal{M}_r}(\boldsymbol{\Theta}^{\star})||_{w_p-w,\star}+||\widehat{\boldsymbol{\Delta}}''||_{w_p-w,\star}+||\Pi_{\overline{\mathcal{M}}_r^{\perp}}(\boldsymbol{\Theta}^{\star})||_{w_p-w,\star}+ \\
&\qquad ||\widehat{\boldsymbol{\Delta}}'||_{w_p-w,\star}\Big\}-\Big\{||\Pi_{\mathcal{M}_r}(\boldsymbol{\Theta}^{\star})||_{w_p-w,\star}+||\Pi_{\overline{\mathcal{M}}_r^{\perp}}(\boldsymbol{\Theta}^{\star})||_{w_p-w,\star}\Big\} \\
&= ||\widehat{\boldsymbol{\Delta}}''||_{w_p-w,\star}+||\widehat{\boldsymbol{\Delta}}'||_{w_p-w,\star}. \qquad (46)
\end{aligned}$$

By combining the inequalities (45) and (46), we can obtain an upper-bound on the Eq. (42);

$$\begin{aligned}
&||\boldsymbol{\Theta}^{\star}||_{w,\star}-||\widehat{\boldsymbol{\Delta}}+\boldsymbol{\Theta}^{\star}||_{w,\star} \\
&\quad = w_p(||\boldsymbol{\Theta}^{\star}||_{\star}-||\widehat{\boldsymbol{\Delta}}+\boldsymbol{\Theta}^{\star}||_{\star})+(||\widehat{\boldsymbol{\Delta}}+\boldsymbol{\Theta}^{\star}||_{w_p-w,\star}-||\boldsymbol{\Theta}^{\star}||_{w_p-w,\star}) \\
&\quad \leq w_p\Big\{2||\Pi_{\overline{\mathcal{M}}_r^{\perp}}(\boldsymbol{\Theta}^{\star})||_{\star}+||\widehat{\boldsymbol{\Delta}}'||_{\star}-||\widehat{\boldsymbol{\Delta}}''||_{\star}\Big\}+\Big\{||\widehat{\boldsymbol{\Delta}}''||_{w_p-w,\star}+||\widehat{\boldsymbol{\Delta}}'||_{w_p-w,\star}\Big\}.
\end{aligned}$$

Now, we control the first term of right-hand side in (41) as follows:

$$\left|\frac{1}{n}\mathbf{tr}(\widehat{\boldsymbol{\Delta}}^{\top}\boldsymbol{X}^{\top}\boldsymbol{E})\right| \leq \left\|\frac{1}{n}\boldsymbol{X}^{\top}\boldsymbol{E}\right\|_{\mathrm{op}}||\widehat{\boldsymbol{\Delta}}||_{\star} \leq \frac{\boldsymbol{\lambda}_n}{2}||\widehat{\boldsymbol{\Delta}}||_{\star}. \qquad (47)$$

In the first inequality, we used Hölder's inequality and in the second inequality the condition $\boldsymbol{\lambda}_n \geq \frac{2}{n}\left\|\boldsymbol{X}^{\top}\boldsymbol{E}\right\|_{\mathrm{op}}$ is used. Combining everything, we finally have a bound on Eq. (41):

$$\begin{aligned}
0 &\leq \frac{1}{n}\mathbf{tr}(\widehat{\boldsymbol{\Delta}}^{\top}\boldsymbol{X}^{\top}\boldsymbol{E})+\boldsymbol{\lambda}_n\Big\{||\boldsymbol{\Theta}^{\star}||_{w,\star}-||\widehat{\boldsymbol{\Delta}}+\boldsymbol{\Theta}^{\star}||_{w,\star}\Big\} \\
&\leq \boldsymbol{\lambda}_n\left\{\frac{1}{2}||\widehat{\boldsymbol{\Delta}}||_{\star}+w_p\Big\{2||\Pi_{\overline{\mathcal{M}}_r^{\perp}}(\boldsymbol{\Theta}^{\star})||_{\star}+||\widehat{\boldsymbol{\Delta}}'||_{\star}-||\widehat{\boldsymbol{\Delta}}''||_{\star}\Big\}+\Big\{||\widehat{\boldsymbol{\Delta}}''||_{w_p-w,\star}+||\widehat{\boldsymbol{\Delta}}'||_{w_p-w,\star}\Big\}\right\} \\
&\leq \boldsymbol{\lambda}_n\left\{\frac{1}{2}||\widehat{\boldsymbol{\Delta}}'||_{\star}+\frac{1}{2}||\widehat{\boldsymbol{\Delta}}''||_{\star}+w_p\Big\{2||\Pi_{\overline{\mathcal{M}}_r^{\perp}}(\boldsymbol{\Theta}^{\star})||_{\star}+||\widehat{\boldsymbol{\Delta}}'||_{\star}-||\widehat{\boldsymbol{\Delta}}''||_{\star}\Big\}+\Big\{||\widehat{\boldsymbol{\Delta}}''||_{w_p-w,\star}+||\widehat{\boldsymbol{\Delta}}'||_{w_p-w,\star}\Big\}\right\} \\
&= \boldsymbol{\lambda}_n\left\{2w_p||\Pi_{\overline{\mathcal{M}}_r^{\perp}}(\boldsymbol{\Theta}^{\star})||_{\star}+||\widehat{\boldsymbol{\Delta}}'||_{2w_p-w+\frac{1}{2},\star}-||\widehat{\boldsymbol{\Delta}}''||_{w-\frac{1}{2},\star}\right\}. \qquad (48)
\end{aligned}$$

Note the norm denotes $||\cdot||_{2w_p-w+\frac{1}{2},\star} := \sum_{j=1}^{p}(2w_p-w_j+\frac{1}{2})\sigma_j(\cdot)$. The inequality (48) implies

$$\sum_{j=1}^{p}\Big(w_j-\frac{1}{2}\Big)\sigma_j(\widehat{\boldsymbol{\Delta}}'') \leq 2w_p\cdot||\Pi_{\overline{\mathcal{M}}_r^{\perp}}(\boldsymbol{\Theta}^{\star})||_{\star}+\sum_{j=1}^{2r}\Big(2w_p-w_j+\frac{1}{2}\Big)\sigma_j(\widehat{\boldsymbol{\Delta}}'). \qquad (49)$$

In (49), we use the fact $\mathrm{rank}(\widehat{\boldsymbol{\Delta}}') \leq 2r$. See the proof of Lemma 1 in Negahban and Wainwright (2011). Because $(w_1-\frac{1}{2})\sum_{j=1}^{p}\sigma_j(\widehat{\boldsymbol{\Delta}}'') \leq \sum_{j=1}^{p}(w_j-\frac{1}{2})\sigma_j(\widehat{\boldsymbol{\Delta}}'')$, and similarly, $\sum_{j=1}^{2r}(2w_p-w_j+\frac{1}{2})\sigma_j(\widehat{\boldsymbol{\Delta}}') \leq (2w_p-w_1+\frac{1}{2})\sum_{j=1}^{2r}\sigma_j(\widehat{\boldsymbol{\Delta}}')$, the inequality (49) implies

$$||\widehat{\boldsymbol{\Delta}}''||_{\star} \leq \frac{2w_p}{w_1-\frac{1}{2}}\sum_{j=r+1}^{p}\sigma_j(\boldsymbol{\Theta}^{\star})+\frac{2w_p-w_1+\frac{1}{2}}{w_1-\frac{1}{2}}\cdot||\widehat{\boldsymbol{\Delta}}'||_{\star}. \qquad (50)$$

8.5 | Proof of Theorem 2

First, recall the basic inequality (40), transformation of weighted nuclear norm (42) and duality of operator and nuclear norm (47). Then, we have

$$\begin{aligned}
\frac{1}{2n}\left\|\boldsymbol{X}\widehat{\boldsymbol{\Delta}}\right\|_{\mathsf{F}}^2 &\leq \frac{1}{n}\mathbf{tr}(\widehat{\boldsymbol{\Delta}}^\top \boldsymbol{X}^\top \boldsymbol{E}) + \boldsymbol{\lambda}_n(||\boldsymbol{\Theta}^\star||_{w,\star} - ||\widehat{\boldsymbol{\Theta}}||_{w,\star}) \\
&\leq \boldsymbol{\lambda}_n\Big\{\frac{1}{2}||\widehat{\boldsymbol{\Delta}}'||_\star + \frac{1}{2}||\widehat{\boldsymbol{\Delta}}''||_\star + w_p(||\boldsymbol{\Theta}^\star||_\star - ||\widehat{\boldsymbol{\Delta}} + \boldsymbol{\Theta}^\star||_\star) \\
&\qquad\qquad + (||\widehat{\boldsymbol{\Delta}} + \boldsymbol{\Theta}^\star||_{w_p - w,\star} - ||\boldsymbol{\Theta}^\star||_{w_p - w,\star})\Big\} \\
&\leq \boldsymbol{\lambda}_n\Big\{\frac{1}{2}||\widehat{\boldsymbol{\Delta}}'||_\star + \frac{1}{2}||\widehat{\boldsymbol{\Delta}}''||_\star + w_p(||\widehat{\boldsymbol{\Delta}}'||_\star + ||\widehat{\boldsymbol{\Delta}}''||_\star) \\
&\qquad\qquad + (||\widehat{\boldsymbol{\Delta}}'||_{w_p - w,\star} + ||\widehat{\boldsymbol{\Delta}}''||_{w_p - w,\star})\Big\} \\
&= \boldsymbol{\lambda}_n\Big\{\sum_{j=1}^{2r}\Big(2w_p - w_j + \frac{1}{2}\Big)\sigma_j(\widehat{\boldsymbol{\Delta}}') + \sum_{j=1}^{p}\Big(2w_p - w_j + \frac{1}{2}\Big)\sigma_j(\widehat{\boldsymbol{\Delta}}'')\Big\} \\
&\leq \boldsymbol{\lambda}_n\Big(2w_p - w_1 + \frac{1}{2}\Big)\Big(\left\|\widehat{\boldsymbol{\Delta}}'\right\|_\star + \left\|\widehat{\boldsymbol{\Delta}}''\right\|_\star\Big), \qquad (51)
\end{aligned}$$

where in the third inequality, triangle inequality of norms $\|\cdot\|_\star$ and $\|\cdot\|_{w_p-w,\star}$ is applied twice. In the last inequality, we used $\sum_{j=1}^{2r}(2w_p - w_j + \frac{1}{2})\sigma_j(\widehat{\boldsymbol{\Delta}}') \leq (2w_p - w_1 + \frac{1}{2})\sum_{j=1}^{2r}\sigma_j(\widehat{\boldsymbol{\Delta}}')$ and $\sum_{j=1}^{p}(2w_p - w_j + \frac{1}{2})\sigma_j(\widehat{\boldsymbol{\Delta}}'') \leq (2w_p - w_1 + \frac{1}{2})\sum_{j=1}^{p}\sigma_j(\widehat{\boldsymbol{\Delta}}'')$.

By the RSC condition, there exists a constant $\kappa > 0$ such that $\kappa\|\widehat{\boldsymbol{\Delta}}\|_{\mathsf{F}}^2 \leq \frac{1}{2n}\left\|\boldsymbol{X}\widehat{\boldsymbol{\Delta}}\right\|_{\mathsf{F}}^2$. Then, by (49) and (51), with some straightforward calculations, we have

$$\begin{aligned}
\kappa\|\widehat{\boldsymbol{\Delta}}\|_{\mathsf{F}}^2 &\leq \boldsymbol{\lambda}_n\frac{2w_p(2w_p - w_1 + \frac{1}{2})}{w_1 - \frac{1}{2}}\cdot\Big(\left\|\widehat{\boldsymbol{\Delta}}'\right\|_\star + \sum_{j=r+1}^{p}\sigma_j(\boldsymbol{\Theta}^\star)\Big) \\
&\leq \boldsymbol{\lambda}_n\frac{2w_p(2w_p - w_1 + \frac{1}{2})}{w_1 - \frac{1}{2}}\cdot\Big(2\sqrt{r}\left\|\widehat{\boldsymbol{\Delta}}\right\|_{\mathsf{F}} + \sum_{j=r+1}^{p}\sigma_j(\boldsymbol{\Theta}^\star)\Big) \\
&\leq \boldsymbol{\lambda}_n\frac{w_p(2w_p - w_1 + \frac{1}{2})}{w_1 - \frac{1}{2}}\cdot\max\Big\{8\sqrt{r}\left\|\widehat{\boldsymbol{\Delta}}\right\|_{\mathsf{F}}, 4\sum_{j=r+1}^{p}\sigma_j(\boldsymbol{\Theta}^\star)\Big\},
\end{aligned}$$

where in the second inequality, we used the fact $\|\widehat{\boldsymbol{\Delta}}'\|_\star \leq \sqrt{2r}\|\widehat{\boldsymbol{\Delta}}'\|_{\mathsf{F}} \leq 2\sqrt{r}\|\widehat{\boldsymbol{\Delta}}\|_F$, and in the last inequality, the inequality $a + b \leq \max\{2a, 2b\}$ for $a, b \geq 0$ is used. Let us denote $\mathcal{W} := \frac{w_p(2w_p - w_1 + \frac{1}{2})}{w_1 - \frac{1}{2}}$. Then, we obtain the final bound:

$$\left\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^\star\right\|_{\mathsf{F}} \leq \max\Big\{8\mathcal{W}\cdot\frac{\boldsymbol{\lambda}_n\sqrt{r}}{\kappa}, \Big[4\mathcal{W}\cdot\frac{\boldsymbol{\lambda}_n\sum_{j=r+1}^{p}\sigma_j(\boldsymbol{\Theta}^\star)}{\kappa}\Big]^{1/2}\Big\}. \qquad (52)$$

Let us construct a set of indices whose corresponding eigenvalues are greater than a threshold $\tau > 0$, and denote it as $\mathcal{K}$ and its complement as $\mathcal{K}^c$.

$$\mathcal{K} := \Big\{j \in \{1, \ldots, p\} : \sigma_j(\boldsymbol{\Theta}^\star) > \tau\Big\}, \qquad \mathcal{K}^c := \Big\{j \in \{1, \ldots, p\} : \sigma_j(\boldsymbol{\Theta}^\star) \leq \tau\Big\}.$$

Since it is assumed that $\boldsymbol{\Theta}^\star \in \mathbb{B}_q(r^\star)$, for $q \in [0,1]$, we have a following inequality:

$$r^\star \geq \sum_{j=1}^{p} \left|\sigma_j(\boldsymbol{\Theta}^\star)\right|^q \geq |\mathcal{K}| \cdot \tau^q, \tag{53}$$

where $|\mathcal{K}|$ denotes a cardinality of the set $\mathcal{K}$. Similarly, by using the definition of set $\mathcal{K}^c$, for $q \in [0,1]$, we have a following inequality:

$$\sum_{j\in\mathcal{K}^c} \sigma_j(\boldsymbol{\Theta}^\star) = \tau \sum_{j=|\mathcal{K}|+1}^{p} \frac{\sigma_j(\boldsymbol{\Theta}^\star)}{\tau} \leq \tau \sum_{j=|\mathcal{K}|+1}^{p} \left(\frac{\sigma_j(\boldsymbol{\Theta}^\star)}{\tau}\right)^q \leq r^\star \cdot (\tau^{1-q}). \tag{54}$$

Set $r = |\mathcal{K}|$ and plugging (53) and (54) in (52) yields:

$$\left\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^\star\right\|_{\mathrm{F}} \leq \max\left\{8\mathcal{W} \cdot \frac{\boldsymbol{\lambda}_n \sqrt{r^\star} \cdot (\tau^{-q/2})}{\kappa}, \left[4\mathcal{W} \cdot \frac{\boldsymbol{\lambda}_n r^\star \cdot (\tau^{1-q})}{\kappa}\right]^{1/2}\right\}. \tag{55}$$

Setting $\tau = \boldsymbol{\lambda}_n/\kappa$ yields that

$$\left\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^\star\right\|_{\mathrm{F}} \leq 8\mathcal{W} \cdot \sqrt{r^\star}\left(\frac{\boldsymbol{\lambda}_n}{\kappa}\right)^{1-q/2}. \tag{56}$$

Recall that we choose $\boldsymbol{\lambda}_n$ such that $\boldsymbol{\lambda}_n \geq \frac{2}{n}\left\|\boldsymbol{X}^\top \boldsymbol{E}\right\|_{\mathrm{op}}$. Negahban and Wainwright (2011) proved that $\frac{2}{n}\left\|\boldsymbol{X}^\top \boldsymbol{E}\right\|_{\mathrm{op}} \leq 10\sigma\|\boldsymbol{\Sigma}\|_{\mathrm{op}}\sqrt{\frac{d_1+d_2}{n}}$ holds with high probability in Lemma 3 of their paper. We formally re-state the Lemma in the following.

Lemma 9 *There are universal constants $c_1, c_2 > 0$ such that*

$$\mathbb{P}\left\{\left|\frac{1}{n}\left\|\boldsymbol{X}^\top \boldsymbol{E}\right\|_{op}\right| \geq 5\sigma\|\boldsymbol{\Sigma}\|_{op}\sqrt{\frac{d_1+d_2}{n}}\right\} \leq c_1 \exp\Big(-c_2(d_1+d_2)\Big).$$

Then, it remains us to determine the constant term κ, which satisfies the RSC property. Readers can refer Lemma 2 in Negahban and Wainwright (2011) for the following result.

Lemma 10 *Let $\boldsymbol{X} \in \mathbb{R}^{n\times d_1}$ be a random matrix with i.i.d. rows sampled from a d_1- variate $\mathcal{N}(0,\Sigma)$ distribution. Then for $n \geq 2d_1$, we have*

$$\mathbb{P}\left\{\sigma_{min}\left(\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}\right) \geq \frac{\sigma_{min}(\boldsymbol{\Sigma})}{9}\right\} \geq 1 - 4\exp\Big(-\frac{n}{2}\Big).$$

With the result from Lemma 10, some algebra shows that we can easily establish the lower bound on the quantity $\frac{1}{2n}\left\|\boldsymbol{X}\widehat{\Delta}\right\|_{\mathrm{F}}^2$ as follows:

$$\frac{1}{2n}\left\|\boldsymbol{X}\widehat{\Delta}\right\|_{\mathrm{F}}^2 = \frac{1}{2n}\sum_{j=1}^{d_2}\left\|(\boldsymbol{X}\widehat{\Delta})_j\right\|_2^2 \geq \frac{1}{2n}\sigma_{\min}(\boldsymbol{X}^\top\boldsymbol{X})\left\|\widehat{\Delta}\right\|_{\mathrm{F}}^2 \geq \frac{\sigma_{\min}(\boldsymbol{\Sigma})}{18}\left\|\widehat{\Delta}\right\|_{\mathrm{F}}^2.$$

This shows that the RSC property holds with probability at least $1 - 4\exp(-n/2)$ with the constant $\kappa = \frac{\sigma_{\min}(\boldsymbol{\Sigma})}{18}$.

Plugging $\boldsymbol{\lambda}_n = 10\sigma \|\boldsymbol{\Sigma}\|_{\text{op}} \sqrt{\frac{d_1+d_2}{n}}$ and $\kappa = \frac{\sigma_{\min}(\boldsymbol{\Sigma})}{18}$ in (56) yields a following inequality:

$$\begin{aligned} \left\| \widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^\star \right\|_{\mathrm{F}}^2 &\leq 64 \mathcal{W}^2 \cdot r^\star \left(10\sigma \|\boldsymbol{\Sigma}\|_{\text{op}} \sqrt{\frac{d_1 + d_2}{n}} \frac{18}{\sigma_{\min}(\boldsymbol{\Sigma})} \right)^{2-q} \\ &= c_1 \mathcal{W}^2 \left(\frac{\sigma^2 \|\boldsymbol{\Sigma}\|_{\text{op}}^2}{\sigma_{\min}^2(\boldsymbol{\Sigma})} \right)^{1-q/2} r^\star \left(\frac{d_1 + d_2}{n} \right)^{1-q/2}. \end{aligned}$$