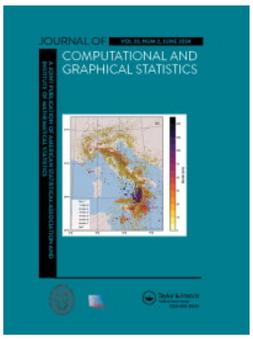

Taylor & Francis
Taylor & Francis Group

Journal of Computational and Graphical Statistics

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/ucgs20

High-Dimensional Multivariate Linear Regression with Weighted Nuclear Norm Regularization

Namjoon Suh, Li-Hsiang Lin & Xiaoming Huo

To cite this article: Namjoon Suh, Li-Hsiang Lin & Xiaoming Huo (26 Apr 2024): High-Dimensional Multivariate Linear Regression with Weighted Nuclear Norm Regularization, Journal of Computational and Graphical Statistics, DOI: 10.1080/10618600.2024.2331020

To link to this article: https://doi.org/10.1080/10618600.2024.2331020

JOURNAL OF COMPUTATIONAL AND GRAPHICAL STATISTICS
2024, VOL. 00, NO. 0, 1–12
https://doi.org/10.1080/10618600.2024.2331020

High-Dimensional Multivariate Linear Regression with Weighted Nuclear Norm Regularization

Namjoon Suh[a], Li-Hsiang Lin[b], and Xiaoming Huo[c]

[a]Department of Statistics and Data Science, UCLA, Los Angeles, CA; [b]Department of Mathematics and Statistics, Georgia State University, Atlanta, GA; [c]H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA

ABSTRACT

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) penalty defined as the weighted sum of the singular values of the matrix. The weights are set in a nondecreasing order, which yields the non-convexity of the WNN objective function in the parameter space. Although the objective function has been widely applied, studies on the estimation properties of its resulting estimator are limited. We propose an efficient algorithm under the framework of the alternative directional method of multipliers (ADMM) to estimate the coefficient matrix. The estimator from the suggested algorithm converges to a stationary point of an augmented Lagrangian function. Under the orthogonal design setting, the effects of the weights for estimating the singular values of the ground-truth coefficient matrix are derived. Under the Gaussian design setting, a minimax convergence rate on the estimation error is derived. We also propose a generalized cross-validation (GCV) criterion for selecting the tuning parameter and an iterative algorithm for updating the weights. Simulations and a real data analysis demonstrate the competitive performance of our new method. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received May 2022
Accepted March 2024

KEYWORDS
Generalized cross-validation;
Low-rank matrix;
Non-convex optimization;
Weighted nuclear norm

1. Introduction

We consider the problem of recovering an unknown coefficient matrix $\boldsymbol{\Theta}^\star \in \boldsymbol{R}^{d_1 \times d_2}$ from n observations of the response vector $y_i \in \mathbb{R}^{d_2}$, $1 \le i \le n$, and predictor $x_i \in \mathbb{R}^{d_1}$, where the ground truth model is as follows:

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

where $\boldsymbol{Y} = (y_1, \ldots, y_n)^\top$ 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 article, we write $p := \min(d_1, d_2)$, $r^\star := \operatorname{rank}(\boldsymbol{\Theta}^\star)$ 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 attractive especially when a dependence structure exists 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 $\boldsymbol{\Theta}^\star$ has a low rank, that is $r^\star \ll p$.

Given the noisy measurement pair $(\boldsymbol{X}, \boldsymbol{Y})$, estimating the ground-truth $\boldsymbol{\Theta}^\star$ with the consistent rank has been intensively studied by many researchers during the past decades. Among them, Yuan et al. (2007) suggested the least-squares 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, the estimator from the standard nuclear norm (SNN) penalized least-squares method still suffers from the bias introduced by the penalization and generally has a higher rank estimate than other methods. To mitigate this issue, Chen, Dong, and Chan (2013) examined 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, Dong, and Chan (2013) considered the WNN penalization on $\boldsymbol{X}\boldsymbol{\Theta}$ instead of directly on $\boldsymbol{\Theta}$, where $\boldsymbol{\Theta}$ is the parameter of interest for inference.

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

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

with the weighted nuclear norm

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

CONTACT Li-Hsiang Lin lhlin@gsu.edu Department of Mathematics and Statistics, Georgia State University, Atlanta, GA 30302.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JCGS.

where $\sigma_j(\boldsymbol{\Theta})$ means the jth largest singular value of a matrix $\boldsymbol{\Theta} \in \boldsymbol{R}^{d_1 \times d_2}$, $\boldsymbol{w} = (w_1, \ldots, w_p)$, w_j is a nonnegative weight assigned to $\sigma_j(\boldsymbol{\Theta})$, $\lambda_n \geq 0$ is a hyper-parameter, and $\|\cdot\|_{\mathrm{F}} := \sqrt{\sum_{j=1}^{p} \sigma_j(\cdot)^2}$ denotes the Frobenius norm. It is a well-known fact that the landscape of (2) is non-convex when the weights are in a non-decreasing order: that is, $0 \leq w_1 \leq w_2 \leq \cdots \leq w_p$. Under this setting, the non-convexity of (2) arises from the violation of triangle inequality of WNN (see Section 2 of Chen, Dong, and Chan (2013)), and (3) should be understood as a semi-norm. However, for the sake of consistency with established conventions in the literature (Chen, Dong, and Chan 2013; Dong et al. 2014; Gu et al. 2014; Zha et al. 2017; Kim, Cho, and Kang 2020) we choose to refer it as a norm. Hereafter, our article only considers the case of nondecreasing weights.

1.1. Contributions

We apply the classical alternative direction method of multipliers (ADMM) algorithm (Boyd, Parikh, and Chu 2011) to solve problem (2) and show that the sequence of tuples generated from the suggested algorithm converges to a stationary point of the augmented Lagrangian function. We refer to our algorithm as WMVR-ADMM where the WMVR stands for Weighted Multi-Variate Regression. This should be contrasted with the result from Chen, Dong, and Chan (2013), in which they provided the closed-form solution of the $\widehat{\boldsymbol{\Theta}}$ to (2), not with the penalization $\|\boldsymbol{\Theta}\|_{\boldsymbol{w},\star}$ but with $\|\boldsymbol{X}\boldsymbol{\Theta}\|_{\boldsymbol{w},\star}$ (see Corollary 1 in their paper). Furthermore, the theoretical analysis of Chen, Dong, and Chan (2013) is focused on the behavior of prediction error, not the estimation error, which is one of the key theoretical findings in our paper.

Our article provides a theoretical explanation of 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 global minimizer of (2) denoted by $\widehat{\boldsymbol{\Theta}}^{\mathrm{OR}}$ and provide a non-asymptotic convergence rate of its singular values to its ground-truth counterparts. Furthermore, we show that for the estimation of nonzero singular values (i.e., $\sigma_j(\boldsymbol{\Theta}^\star) > 0$), setting small weights w_j (i.e., $w_j < 2\sigma$) is desirable. For zero singular values (i.e., $\sigma_j(\boldsymbol{\Theta}^\star) = 0$), large weights (i.e., $w_j > 2\sigma$) are required for achieving fast convergence rates to the ground-truth. 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 the high-dimensional regime (i.e., $n \ll d_1 d_2$).

Finally, we develop a data-driven method for choosing the value of the tuning parameters in the model. For updating the weights (w_j's) on the singular value, we borrow the idea from the seminal work of Candes, Wakin, and Boyd (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 λ_n, we adopt a generalized cross-validation (GCV) type of criterion. This is enabled through the development of a surrogate function (13), 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, which makes the GCV statistic computable.

The following example demonstrates the advantage of our proposed method for estimating the singular values of $\boldsymbol{\Theta}^\star$ when it is compared with the traditional SNN method. We consider a setting of coefficient matrix $\boldsymbol{\Theta}^\star \in \mathbb{R}^{250\times 250}$ with $r^\star = 50$ and generate $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{250\times 50}$ with each entry from $\mathcal{N}(0,1)$ and set $\boldsymbol{\Theta}^\star = \boldsymbol{A}\boldsymbol{B}^\top$. Each entry of $\boldsymbol{X} \in \mathbb{R}^{n\times d_1}$ is sampled from $\mathcal{N}(0,1)$. Variance parameter σ^2 is set as 1, and the hyper-tuning parameter λ_n is set at $5\sqrt{\frac{d_1+d_2}{n}}$, where $d_1 = d_2 = 250$. Figure 1 displays 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) of Figure 1 exhibits the result of the first iteration of WMVR-ADMM with sample size $n = 250$. Note that we start the WMVR-ADMM with $\{w_j\}_{j=1}^{p} = 1$, equivalent to solving SNN problem. Panel (B) presents the second iteration results of the algorithm with the updated weights based on the weight update rule presented in Section 4.1. Panel (C) displays the result of the SNN problem with $n = 1000$. Notice that WMVR-ADMM achieves a satisfactory result within two iterations of the loop (Panels (A) and (B)) with only $n = 250$. In contrast, a slight bias still exists on each of the estimated singular values from SNN with $n = 1000$.

1.2. Additional Related Literature

In the field of computer vision, many papers, including Gu et al. (2014), Gu et al. (2017), Xu et al. (2017), Yair and Michaeli (2018), Liu et al. (2018), and Kim, Cho, and Kang (2020) studied WNN minimization problem in the context of matrix completion. In the statistical literature, we are not aware of many applications of the WNN in matrix regression problems except Chen, Dong, and Chan (2013).

In contrast, there are a myriad of papers that studied the statistical properties of SNN penalized least squares problem under even a more general model than the multivariate linear regression. We only mention a subset of them. Bach (2008) provided necessary and sufficient conditions for the asymptotic rank consistency of the SNN problem, and later Lee, Sun, and Taylor (2015) proved the non-asymptotic rank consistency of the estimator from the SNN under the irrepresentable assumption on the design matrix. Under the sub-Gaussian noise assumption, Negahban and Wainwright (2011) derived a minimax optimal rate of the estimation error of a trace regression model in which $\boldsymbol{\Theta}^\star$ 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, Lounici, and Tsybakov (2011) established a sharp oracle inequality of the trace regression estimator under the restricted isometry condition of the design matrix $\boldsymbol{X}$. In the subsequent work, Fan, Gong, and Zhu (2019) investigated the SNN problem under generalized trace regression problems for the categorical responses. Recently, Fan, Wang, and Zhu (2021) worked on obtaining the same minimax estimation rate of a trace regression problem with Negahban and Wainwright (2011) under the heavy-tail assumption on the design matrix and observational noise.

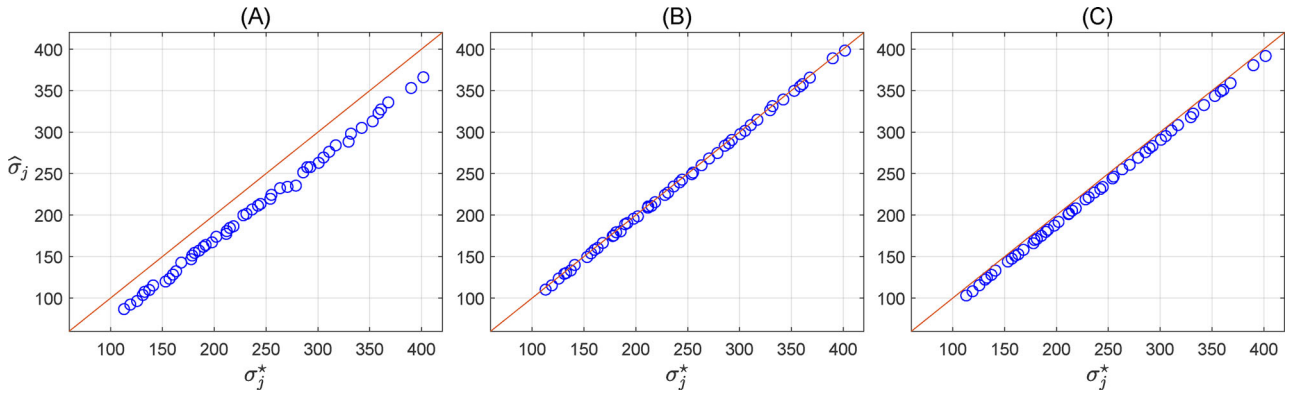

Figure 1. Three panels display the plots of estimated singular values versus the ground truth singular values $\sigma_j^\star$. The first two panels (*A*) and (*B*) are the results from the WMVR-ADMM algorithm of the first and second iterations with one weight update, respectively, under $n = 250$. The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

Our work also falls into the category of the adaptive penalized estimation problem. Among a plethora of papers, 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 squares fit with the adaptive ℓ_1-penalization is always a convex optimization problem. Later, Candes, Wakin, and Boyd (2008) suggested an algorithm for updating the weights in the adaptive lasso algorithm. The main idea of their article is to simply update the weights as the inverse of the estimated coefficient in the previous iteration.

1.3. Organization

The rest of the article is organized as follows. In Section 2, we introduce the details of WMVR-ADMM and provide a theorem, assuring the convergence of the proposed algorithm. In Section 3, the 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 setting, we obtain the minimax rate of the estimation error. In Section 4, a two-stage data-driven method for updating weights and tuning the regularization parameter through GCV statistics is detailed. In Section 5, numerical experiments from both synthetic and real datasets are presented. Specifically, in Section 5.1, our simulation results corroborate the statement in Theorem 2.2. In Section 5.2, under specific simulated scenarios, the performance of WNN estimators is compared with estimators from the SNN method (Yuan et al. 2007), Adaptive Nuclear Norm, in short ANN (Chen, Dong, and Chan 2013), and Reduced Ridge Rank Regression (Mukherjee and Zhu 2011) under two metrics: estimation error and estimated rank. In Section 5.3, our proposed WMVR-ADMM estimator is applied to a real dataset showing the effectiveness of our method. Finally, in Section 6, we conclude our article with a discussion section.

2. WMVR-ADMM and Convergence Guarantee

To develop an algorithm for solving (2), we start with reformulating (2) as follows:

$$\min_{\boldsymbol{\Theta},\boldsymbol{\Gamma}} \left\{ f(\boldsymbol{\Theta}) + g(\boldsymbol{\Gamma}) \right\} \quad \textbf{s.t.} \quad \boldsymbol{\Theta} = \boldsymbol{\Gamma} \in \mathbb{R}^{d_1 \times d_2}, \tag{4}$$

This reformulation leads to the construction of an augmented Lagrangian function $\mathcal{L}_\rho(\boldsymbol{\Theta}, \boldsymbol{\Gamma}, \boldsymbol{\Lambda})$: For any $\rho > 0$ and dual variable $\boldsymbol{\Lambda} \in \mathbb{R}^{d_1 \times d_2}$, we define,

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

Then, we update the estimators through the following three optimization steps iteratively until primal and dual feasibility conditions hold; to be more specific, we repeat Steps 1–3

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

until $\|\boldsymbol{\Theta}^{(k+1)} - \boldsymbol{\Gamma}^{(k+1)}\|_{\mathrm{F}} \le 10^{-5}$ and $\|\boldsymbol{\Gamma}^{(k+1)} - \boldsymbol{\Gamma}^{(k)}\|_{\mathrm{F}} \le 10^{-5}$. Here, we denote the tuple $(\boldsymbol{\Theta}^{(k)}, \boldsymbol{\Gamma}^{(k)}, \boldsymbol{\Lambda}^{(k)})$ as the updated parameters at the kth 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., $\|\cdot\|_{w,\star}$) over $\boldsymbol{\Theta}$ with fixed $\boldsymbol{\Gamma}^{(k)}$ and $\boldsymbol{\Lambda}^{(k)}$, whereas the objective function in Step 2 is a simple quadratic function of $\boldsymbol{\Gamma}$ with fixed $\boldsymbol{\Theta}^{(k+1)}$ and $\boldsymbol{\Lambda}^{(k)}$. The algorithm is conducted by initializing $\boldsymbol{\Theta}^{(0)} = \boldsymbol{\Gamma}^{(0)} = \boldsymbol{\Lambda}^{(0)} = \mathbf{0} \in \mathbb{R}^{d_1 \times d_2}$. Next, the key of our algorithm is that a closed-form solution of Step 1 can be obtained, even though it is a non-convex problem. We state the result in Lemma 2.1 whose proof is deferred in Section A of supplemental material.

Lemma 2.1. Let $\boldsymbol{\Theta}^{(k+1)}$ be the minimizer of Step 1 in (k+1)th iteration. Denote $\boldsymbol{B}^{(k)} := -\boldsymbol{\Lambda}^{(k)} + \rho \cdot \boldsymbol{\Gamma}^{(k)}$ and its singular value decomposition as $\boldsymbol{U}^{\mathbf{B}}\boldsymbol{D}^{\mathbf{B}}(\boldsymbol{V}^{\mathbf{B}})^\top$. Then, for any fixed $\lambda_n, \rho \ge 0$ and $0 \le w_1 \le \cdots \le w_p$,

$$\boldsymbol{\Theta}^{(k+1)} = \boldsymbol{U}^{\mathbf{B}}\mathcal{S}_{\lambda_n w}(\boldsymbol{D}^{\mathbf{B}})(\boldsymbol{V}^{\mathbf{B}})^\top,$$
$$\mathcal{S}_{\lambda_n w}(\boldsymbol{D}^{\mathbf{B}}) = \mathrm{diag}\left\{\max\left\{\frac{1}{\rho}\big(\sigma_j(\boldsymbol{B}^{(k)}) - \lambda_n w_j\big), 0\right\}, j = 1, \ldots, p\right\}.$$

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

For the optimization problem in Step 2, it can be rewritten and solved as follows:

$$\begin{aligned}\boldsymbol{\Gamma}^{(k+1)} &= \underset{\boldsymbol{\Gamma}\in\mathbb{R}^{d_1\times d_2}}{\operatorname{argmin}}\ \mathcal{L}_\rho\big(\boldsymbol{\Theta}^{(k+1)},\boldsymbol{\Gamma},\boldsymbol{\Lambda}^{(k)}\big)\\ &= \underset{\boldsymbol{\Gamma}\in\mathbb{R}^{d_1\times d_2}}{\operatorname{argmin}}\ \left\{\operatorname{tr}\left(\boldsymbol{\Gamma}^\top\left(\frac{1}{2n}\boldsymbol{X}^\top\boldsymbol{X}+\frac{\rho}{2}\cdot\mathcal{I}_{d_1\times d_1}\right)\right.\right.\\ &\qquad\left.\left.\boldsymbol{\Gamma}-\left(\frac{1}{n}\boldsymbol{Y}^\top\boldsymbol{X}+\rho\cdot\boldsymbol{\Theta}^{(k+1)}+\boldsymbol{\Lambda}^{(k)}\right)^\top\boldsymbol{\Gamma}\right)\right\} \quad (6)\\ &= \left(\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}+\rho\cdot\mathcal{I}_{d_1\times d_1}\right)^{-1}\\ &\qquad\left(\frac{1}{n}\boldsymbol{Y}^\top\boldsymbol{X}+\rho\cdot\boldsymbol{\Theta}^{(k+1)}+\boldsymbol{\Lambda}^{(k)}\right). \quad (7)\end{aligned}$$

Note that the quadratic equation (6) always has a unique minimizer (7) as long as $\rho > 0$. With the updated $\boldsymbol{\Theta}^{(k+1)}$ and $\boldsymbol{\Gamma}^{(k+1)}$ from Steps 1, 2, we can easily update $\boldsymbol{\Lambda}^{(k)}$ to $\boldsymbol{\Lambda}^{(k+1)}$ through Step 3. The final output of WMVR-ADMM is a minimizer of $\mathcal{L}_\rho\big(\boldsymbol{\Theta},\boldsymbol{\Gamma}^{(\mathcal{T}-1)},\boldsymbol{\Lambda}^{(\mathcal{T}-1)}\big)$ in Step 1, where $\mathcal{T}$ denotes the last iteration index of the algorithm. The implementation is summarized in Algorithm 1. Note that the WMVR-ADMM algorithm can be easily extended to the trace regression model, which is a generalization of the multivariate linear regression model.[1] The convergence of the WMVR-ADMM is shown in Theorem 2.2 with its proof given in Section B of supplemental material. The proof is motivated from Wang, Yin, and Zeng (2019) and Kim, Cho, and Kang (2020).

Theorem 2.2. Set $\rho > 2L_{\nabla g}$ with $L_{\nabla g} := \sigma_1\big(\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}\big)$. The sequence $\{(\boldsymbol{\Theta}^{(k)},\boldsymbol{\Gamma}^{(k)},\boldsymbol{\Lambda}^{(k)})\}_{k\geq 1}$ from Algorithm 1 converges to a limit point $(\boldsymbol{\Theta}^\star,\boldsymbol{\Gamma}^\star,\boldsymbol{\Lambda}^\star)$ regardless of initialized tuple $(\boldsymbol{\Theta}^{(0)},\boldsymbol{\Gamma}^{(0)},\boldsymbol{\Lambda}^{(0)})$. The limit point $(\boldsymbol{\Theta}^\star,\boldsymbol{\Gamma}^\star,\boldsymbol{\Lambda}^\star)$ is a stationary point of $\mathcal{L}_\rho$.

The threshold for penalty parameter ρ (i.e., $L_{\nabla g} := \sigma_1\big(\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}\big)$) can be computed from data. The theorem states that the sequence generated by WMVR-ADMM converges to a certain stationary point regardless of the initialized tuple $(\boldsymbol{\Theta}^{(0)},\boldsymbol{\Gamma}^{(0)},\boldsymbol{\Lambda}^{(0)})$. The statement is corroborated by a set of numerical experiments in Section 5.1 which shows that the sequences $\{(\boldsymbol{\Theta}^{(k)},\boldsymbol{\Gamma}^{(k)},\boldsymbol{\Lambda}^{(k)})\}_{k\geq 1}$ with varying initialized tuples converge to the solution with the same objective value. Wang, Yin, and Zeng (2019) named the convergence of tuples to the certain limit point $(\boldsymbol{\Theta}^\star,\boldsymbol{\Gamma}^\star,\boldsymbol{\Lambda}^\star)$ as "global" convergence in a sense that it is not affected by the initialized tuple. This global convergence comes as a consequence of (5) being a Kurdyka-Łojasiewicz (KL) function. See Proposition 2 in Wang, Yin, and Zeng (2019) and Theorem 2.9 in Attouch, Bolte, and Svaiter (2013) for this claim.

[1] Refer Negahban and Wainwright (2011) for checking how to translate MVLR to trace regression model. A description of the extended algorithm of WMVR-ADMM to trace regression is provided in Section G of supplemental material.

Input : A measurement pair $(\boldsymbol{X},\boldsymbol{Y})$, $\lambda_n \geq 0$ and weights $0 \leq w_1 \leq \cdots \leq w_p$.
Initialization : $\boldsymbol{\Theta}^{(0)} = \boldsymbol{\Gamma}^{(0)} = \boldsymbol{\Lambda}^{(0)} = \mathbf{0} \in \mathbb{R}^{d_1\times d_2}$.
Repeat the following steps:
Step 1. Let $\boldsymbol{B}^{(k)} := -\boldsymbol{\Lambda}^{(k)} + \rho\cdot\boldsymbol{\Gamma}^{(k)}$.
$\boldsymbol{B}^{(k)} = \boldsymbol{U}^{\mathbf{B}}\boldsymbol{D}^{\mathbf{B}}\big(\boldsymbol{V}^{\mathbf{B}}\big)^\top$.
Set $\mathcal{S}_{\lambda_n w}\big(\boldsymbol{D}^{\mathbf{B}}\big) = \mathbf{diag}\Big\{\max\Big\{\frac{1}{\rho}\big(\sigma_j(\boldsymbol{B}^{(k)}) - \lambda_n w_j\big), 0\Big\}$ for $j = 1,\ldots,p\Big\}$.
$\boldsymbol{\Theta}^{(k+1)} = \boldsymbol{U}^{\mathbf{B}}\mathcal{S}_{\lambda_n w}\big(\boldsymbol{D}^{\mathbf{B}}\big)\big(\boldsymbol{V}^{\mathbf{B}}\big)^\top$
Step 2. $\boldsymbol{\Gamma}^{(k+1)} = \big(\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}+\rho\cdot\mathcal{I}_{d_1\times d_1}\big)^{-1}\big(\frac{1}{n}\boldsymbol{Y}^\top\boldsymbol{X}+\rho\cdot\boldsymbol{\Theta}^{(k+1)}+\boldsymbol{\Lambda}^{(k)}\big)$.
Step 3. $\boldsymbol{\Lambda}^{(k+1)} = \boldsymbol{\Lambda}^{(k)} + \rho\big(\boldsymbol{\Theta}^{(k+1)} - \boldsymbol{\Gamma}^{(k+1)}\big)$.
Until $\|\boldsymbol{\Theta}^{(k+1)} - \boldsymbol{\Gamma}^{(k+1)}\|_{\mathrm{F}} \leq 10^{-5}$ and $\|\boldsymbol{\Gamma}^{(k+1)} - \boldsymbol{\Gamma}^{(k)}\|_{\mathrm{F}} \leq 10^{-5}$.
Output : $\widehat{\boldsymbol{\Theta}} = \boldsymbol{\Theta}^{(k+1)}$.

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

Remark 1. Note that Theorem 2.2 does not necessarily guarantee the convergence of WMVR-ADMM to the global minimizer of (2). From our theoretical result in Theorem 2.2, Algorithm 1 guarantees the convergence of the sequence to a certain limit point $(\boldsymbol{\Theta}^\star,\boldsymbol{\Gamma}^\star,\boldsymbol{\Lambda}^\star)$. Given $\boldsymbol{\Theta}^\star \approx \boldsymbol{\Gamma}^\star$, the Lagrangian function becomes the original objective function (2). Therefore, we can roughly say that the WMVR-ADMM converges to the stationary point of the original function (2). This stationary point can be a saddle point, local minimum, or even global minimum of (2). Although not theoretically justified, under an orthogonal design setting, we empirically observe that the sequence generated from WMVR-ADMM converges to the global optimum of (2) in Section 5.1. This verification is possible since we know the closed-form solution of $\widehat{\boldsymbol{\Theta}}^{\mathrm{OR}}$ which will be given in Proposition 3.1. Further technical comments on this issue are deferred to Section 6.

Remark 2. The closed-form solutions of steps 1 and 2 help us analyze the time complexity of each step in Algorithm 1 easily. The most time-consuming parts are the SVD computation of $\mathbf{B}^{(k)} \in \mathbb{R}^{d_1\times d_2}$, which is $O(d_1^2 d_2 + d_1 d_2^2 + d_2^3)$, and the inverse computation of $(\frac{1}{n}\mathbf{X}^T\mathbf{X} + \rho\mathcal{I}_{d1\times d_1}) \in \mathbb{R}^{d_1\times d_1}$ in step 2, which is $O(d_1^3)$. When $d_1 = d_2$, $p = \min\{d_1, d_2\}$ the complexity of both parts becomes $O(p^3)$. To observe the time cost at various scales of data dimensions, we provide a summary in Section K of supplemental material. The results show that our algorithm can be used for a reasonable time to solve multivariate regression problems with hundreds of data points or variables and is workable under a higher number of variables or a larger sample size. Overall, we see that as the dimension d_1 (or d_2) increases, the running time follows cubic growth as $O(p^3)$ as the theoretical results suggest. Additionally, when the dimension is fixed and the sample size increases, we observe that the running time follows linear growth, mainly contributed by matrix multiplication or linear operations applied to matrices in Algorithm 1.

3. Statistical Properties of the Estimator

3.1. Statistical Properties of $\widehat{\Theta}$ under the Orthogonal Design

We first study the convergence rate of the estimated singular values under the orthogonal design setting, which sheds light on the role of weights in the estimation of singular values.

Proposition 3.1. Let $\widehat{\boldsymbol{U}}^{\mathrm{LS}}\widehat{\boldsymbol{D}}^{\mathrm{LS}}\big(\widehat{\boldsymbol{V}}^{\mathrm{LS}}\big)^{\top}$ be the singular value decomposition of the least-square estimator $\widehat{\boldsymbol{\Theta}}^{\mathrm{LS}} := \big(\boldsymbol{X}^{\top}\boldsymbol{X}\big)^{\dagger}\boldsymbol{X}^{\top}\boldsymbol{Y}$. Then, under the orthogonal design (i.e., $\boldsymbol{X}^{\top}\boldsymbol{X} = n\boldsymbol{I}_{d_1\times d_1}$), the SVD of the minimizer of (2) has the following closed-form solution: $\widehat{\boldsymbol{\Theta}}^{\mathrm{OR}} := \widehat{\boldsymbol{U}}^{\mathrm{LS}}\widehat{\boldsymbol{D}}\big(\widehat{\boldsymbol{V}}^{\mathrm{LS}}\big)^{\top}$, where the diagonal entry of $\widehat{\boldsymbol{D}}$ is $\sigma_j\big(\widehat{\boldsymbol{\Theta}}^{\mathrm{OR}}\big) = \max\big(\sigma_j\big(\widehat{\boldsymbol{\Theta}}^{\mathrm{LS}}\big) - \lambda_n w_j, 0\big)$ for $j = 1,\ldots,p$. Furthermore, suppose $\lambda_n = \sqrt{\frac{d_1+d_2}{n}}$. Then, with probability at least $1 - 2\exp(-(\sqrt{d_1}+\sqrt{d_2})^2/2)$, for j such that $\sigma_j\big(\boldsymbol{\Theta}^{\star}\big) > 0$, we have,

$$\Big|\sigma_j\big(\widehat{\boldsymbol{\Theta}}^{\mathrm{OR}}\big) - \sigma_j\big(\boldsymbol{\Theta}^{\star}\big)\Big| \le \max\big(4\sigma, 2w_j\big)\cdot\sqrt{\frac{d_1+d_2}{n}}. \quad (8)$$

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

$$\Big|\sigma_j\big(\widehat{\boldsymbol{\Theta}}^{\mathrm{OR}}\big)\Big| \le \min\big(2\sigma, w_j\big)\cdot\sqrt{\frac{d_1+d_2}{n}}. \quad (9)$$

The proof of Proposition 3.1 can be found in Section C of supplemental material. Here, $\big(\boldsymbol{X}^{\top}\boldsymbol{X}\big)^{\dagger}$ denotes the Moore-Penrose inverse of enclosed Gram-matrix. Based on the closed-form solution of $\widehat{\boldsymbol{\Theta}}^{\mathrm{OR}}$ in Proposition 3.1, under the orthogonal design assumption, each estimated singular value has a form $\max\big(\sigma_j\big(\widehat{\boldsymbol{\Theta}}^{\mathrm{LS}}\big) - \lambda_n w_j, 0\big)$ for $j \in \{1,\ldots,p\}$. Then, for the fixed λ_n, it is easy to see that the large weights for small singular values of $\widehat{\boldsymbol{\Theta}}^{\mathrm{LS}}$ can induce the sparsity among the singular values of $\widehat{\boldsymbol{\Theta}}^{\mathrm{OR}}$. Furthermore, the proposition states that with an appropriate choice of tuning parameter λ_n, the singular values of the $\widehat{\boldsymbol{\Theta}}^{\mathrm{OR}}$ are consistently estimated. Bounds in (8) and (9) provide us with the guidelines for the choices of weights. For the set of indices $\{j : \sigma_j\big(\boldsymbol{\Theta}^{\star}\big) > 0\}$, the corresponding w_js need to be set lower than twice the magnitude of regression noise, that is, 2σ, whereas, for the set of indices $\{j : \sigma_j\big(\boldsymbol{\Theta}^{\star}\big) = 0\}$, the corresponding weights can be set even higher than 2σ. This is consistent with our intuition that we need small weights for estimating the nonzero singular values of $\boldsymbol{\Theta}^{\star}$, whereas large weights are required for the consistent estimation of zero singular values of $\boldsymbol{\Theta}^{\star}$.

3.2. Estimation Error under Random Design

In this section, we study the estimation error under a random design assumption in the Frobenius norm (i.e., $\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^{\star}\|_{\mathrm{F}}^2$). For the precise statement of the main theorem on the estimation error, two technical assumptions are required : (I) A design matrix $\boldsymbol{X}$ is assumed to be random, whose rows are independently sampled from d_1-variate $\mathcal{N}(0,\boldsymbol{\Sigma})$ distribution for some positive definite covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{d_1\times d_1}$, and (II) The exact low-rank assumption of $\boldsymbol{\Theta}^{\star}$ is relaxed to a nearly low-rank matrix by requiring that the $\{\sigma_j\big(\boldsymbol{\Theta}^{\star}\big)\}_{j=1}^{p}$ decays fast enough. Specifically, for a parameter $q \in (0,1]$ and a radius R_q, we assume that $\boldsymbol{\Theta}^{\star} \in \mathbb{B}_q(R_q) := \big\{\boldsymbol{\Theta} \in \mathbb{R}^{d_1\times d_2} : \sum_{j=1}^{p}\big|\sigma_j\big(\boldsymbol{\Theta}\big)\big|^q \le R_q\big\}$. In a limiting case when $q = 0$, we define the set $\mathbb{B}_0(r^{\star}) := \big\{\boldsymbol{\Theta} \in \mathbb{R}^{d_1\times d_2} : \sum_{j=1}^{p}\mathbb{1}(\sigma_j\big(\boldsymbol{\Theta}\big) \ne 0) \le r^{\star}\big\}$.

3.2.1. Restricted Strong Convexity of Loss function

In this sub-subsection, an important lemma for describing the notion of "restricted strong convexity (RSC)" of the loss function is stated. To be more specific, in high-dimensional setting where $n \ll d_1 d_2$, although the function $\mathcal{L}_n(\boldsymbol{\Theta}) := \frac{1}{2n}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\Theta}\|_{\mathrm{F}}^2$ might be curved in some directions, there are $\big(d_1 d_2 - n\big)$ directions where it is flat up to the second order. We hope that the associated error matrix $\widehat{\boldsymbol{\Delta}} := \widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^{\star}$ lies in some directions $\mathcal{C} \subseteq \mathbb{R}^{d_1\times d_2}$ where the $\mathcal{L}_n(\boldsymbol{\Theta})$ is curved. This notion is expressed as follows: for some positive constant $\kappa > 0$,

$$\mathcal{E}_n\big(\widehat{\boldsymbol{\Delta}}\big) \ge \kappa\|\widehat{\boldsymbol{\Delta}}\|_{\mathrm{F}}^2 \qquad \text{for all} \quad \widehat{\boldsymbol{\Delta}} \in \mathcal{C}, \quad (10)$$

where $\mathcal{E}_n\big(\widehat{\boldsymbol{\Delta}}\big)$ denotes the first order Taylor-expansion error of $\mathcal{L}_n(\cdot)$ around $\boldsymbol{\Theta}^{\star}$.

Before we formally state the lemma that characterizes the set $\mathcal{C}$, let us introduce the relevant notation. Denote $\boldsymbol{U}^{\star}$ and $\boldsymbol{V}^{\star}$ as the left and right singular matrices of $\boldsymbol{\Theta}^{\star}$. Let $r\ (\le p)$ be any arbitrary integer. 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 $\boldsymbol{\Theta}$, respectively. Then, with these notations, the $\mathcal{M}_r(\mathcal{U},\mathcal{V})$ (resp. $\overline{\mathcal{M}}_r^{\perp}(\mathcal{U},\mathcal{V})$) corresponds to a subspace of matrices with nonzero left and right singular vectors associated with the first r (resp. the remaining $(p-r)$) columns of $\boldsymbol{U}^{\star}$ and $\boldsymbol{V}^{\star}$: for any given integer $r\ (\le p)$, we have

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

Hereafter, we will omit $\mathcal{U}$ and $\mathcal{V}$ from the notations, if they are clear from the context. The notations can be used to characterize the set $\mathcal{C}$ as shown in the lemma below:

Lemma 3.2. Suppose $\widehat{\boldsymbol{\Theta}}$ is a global minimizer of (2) with the associated matrix $\widehat{\boldsymbol{\Delta}} = \widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^{\star}$. Set the weights $\frac{1}{2} < w_1 \le \cdots \le w_p$ and suppose regularization parameter is chosen such that $\lambda_n \ge \frac{2}{n}\big\|\boldsymbol{X}^{\top}\boldsymbol{E}\big\|_{\mathrm{op}}$. Let $\|\cdot\|_{\star} := \sum_{j=1}^{p}\sigma_j(\cdot)$. Then, for a positive integer $r \le p$,

$$\begin{aligned}\mathcal{C}(w;r) :=& \Bigg\{\widehat{\boldsymbol{\Delta}} \in \mathbb{R}^{d_1\times d_2} : \|\widehat{\boldsymbol{\Delta}}''\|_{\star} \le \frac{2w_p}{w_1 - \frac{1}{2}}\sum_{j=r+1}^{p}\sigma_j\big(\boldsymbol{\Theta}^{\star}\big)\\ &+ \frac{2w_p - w_1 + \frac{1}{2}}{w_1 - \frac{1}{2}}\cdot\|\widehat{\boldsymbol{\Delta}}'\|_{\star}\Bigg\}, \quad (11)\end{aligned}$$

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

A detailed proof of Lemma 3.2 is deferred in Section D of supplemental material. The result holds for the global minimizer of (2) as the proof relies on the basic inequality, that is, see eq. (16) in supplemental material. The lemma shows that the subset $\mathcal{C}$ corresponds to the matrices $\widehat{\mathbf{\Delta}}$ for which the quantity $\|\widehat{\mathbf{\Delta}}''\|_\star$ is relatively small compared to the weighted sum of $\|\widehat{\mathbf{\Delta}}'\|_\star$ and $\sum_{j=r+1}^{p}\sigma_j(\mathbf{\Theta}^\star)$ with any given $r \le p$. The weights put in $\|\widehat{\mathbf{\Delta}}'\|_\star$ and $\sum_{j=r+1}^{p}\sigma_j(\mathbf{\Theta}^\star)$ are functions of a pair (w_1, w_p), and this pair characterizes the size of the subset $\mathcal{C}$. We restrict the case $w_1 > \frac{1}{2}$ for a technical reason. The closer w_1 gets to $\frac{1}{2}$ and the larger w_p we have, the bigger the size of $\mathcal{C}$ becomes. Also, Lemma 3.2 shows that plugging in $w_1 = \cdots = w_p = 1$ recovers one of the constraints that are used to define the set in Lemma 1 of Negahban and Wainwright (2011).

Remark 3. A notable difference between the set in (11) and the set (12) in Negahban and Wainwright (2011) is the existence of the constraint, $\|\widehat{\mathbf{\Delta}}\|_{\mathrm{F}} \ge \delta$, where $\delta > 0$ is a tolerance parameter. Note that when $\mathrm{rank}(\mathbf{\Theta}^\star) = r$, the set $\mathcal{C}$ becomes a cone. But when $\mathrm{rank}(\mathbf{\Theta}^\star) > r$, it no longer defines a cone where it contains an open ball. The constraint removes certain directions in $\mathbb{R}^{d_1\times d_2}$ on which $\widehat{\Delta}$ can lie so that the RSC condition holds over the set $\mathcal{C}$, even when $\mathcal{E}_n(\widehat{\mathbf{\Delta}})$ fails strong convexity in a global sense. Recall (10). Nonetheless in our setting, as we can ensure $\mathcal{E}_n(\widehat{\mathbf{\Delta}}) \ge \frac{\sigma_{\min}(\mathbf{\Sigma})}{18}\|\widehat{\mathbf{\Delta}}\|_{\mathrm{F}}^2$ for all $\widehat{\Delta} \in \mathbb{R}^{d_1\times d_2}$, where $\sigma_{\min}(\mathbf{\Sigma})$ denotes a minimum eigenvalue of $\mathbf{\Sigma}$, the constraint is not required. We refer readers to the proof of Corollary 3 in Negahban and Wainwright (2011) for this result.

3.2.2. The Main Theorem on Estimation Error

With the RSC condition and Lemma 3.2, we can further show that the estimation error converges to 0 at a minimax rate, whose proof is given in Section E of supplemental material.

Theorem 3.3. The regularization parameter is chosen such that $\lambda_n = 10\sigma\|\mathbf{\Sigma}\|_{\mathrm{op}}\sqrt{\frac{d_1+d_2}{n}}$ and weights are set as $\frac{1}{2} < w_1 \le \cdots \le w_p$. 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 a global minimizer $\widehat{\mathbf{\Theta}}$ of (2) satisfies the following bound:

$$\left\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^\star\right\|_{\mathrm{F}}^2 \le c_1\mathcal{W}^2\left(\frac{\sigma^2\|\mathbf{\Sigma}\|_{\mathrm{op}}^2}{\sigma_{\min}^2(\mathbf{\Sigma})}\right)^{1-q/2}\cdot R_q\left(\frac{d_1+d_2}{n}\right)^{1-q/2}. \tag{12}$$

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

Here, $\|\mathbf{\Sigma}\|_{\mathrm{op}}$ denotes the spectral norm of the matrix $\mathbf{\Sigma}$. Notably, when $\mathbf{\Theta}^\star \in \mathbb{B}_0(r^\star)$ is an exact rank $r^\star$ matrix (i.e., $q = 0$) and $\mathbf{\Sigma} = \mathcal{I}_{d_1\times d_1}$, convergence rate of the estimation error becomes $\mathcal{O}\big(\mathcal{W}^2\frac{\sigma^2 r^\star(d_1+d_2)}{n}\big)$. 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, Lounici, and Tsybakov (2011), and Rohde and Tsybakov (2011). It is worth noting that $\{c_i, i = 1, 2, 3\}$ are the universal constants independent of weights $\{w_j\}_{j=1}^p$ and the information on weights is solely encoded in the factor $\mathcal{W}$. This factor enables a comparison of estimation rates between SNN and WNN. More discussions on this comparison can be found in Section 6.

4. Data-Driven Model Selections

4.1. Weight Update Rule

In this section, we propose an iterative algorithm that alternates between estimating $\mathbf{\Theta}^\star$ and updating weights $\{w_j\}_{j=1}^p$. For any fixed $\boldsymbol{\lambda_n} \ge 0$, we have the following procedure (I) $\sim$ (IV):

(I) Set the iteration count ℓ to 1 and weights $w_1^{(\ell)} = \cdots = w_p^{(\ell)} = 1$.

(II) Solve (2) via WMVR-ADMM with the weights $\{w_j^{(\ell)}\}_{j=1}^p$, and denote the solution as $\widehat{\mathbf{\Theta}}^{(\ell)}$.

(III) For $j \in \{1, \ldots, p\}$, update the weights for the next iteration as $w_j^{(\ell+1)} = (\sigma_j(\widehat{\mathbf{\Theta}}^{(\ell)}) + \varepsilon)^{-1}$ and set $\ell + 1 = \ell$

(IV) Repeat steps (II) and (III) until the following holds: $\mathrm{rank}(\widehat{\mathbf{\Theta}}^{(\ell)}) = \mathrm{rank}(\widehat{\mathbf{\Theta}}^{(\ell+1)})$.

In steps (I) and (II) with $\ell = 1$, we start the algorithm by solving the SNN problem with weights set as $w_j = 1$ for $j = 1, \ldots, p$. In step (II) with $\ell \ge 1$, we employ WMVR-ADMM developed in Section 2. In step (III), weights $\{w_j^{(\ell+1)}\}_{j=1}^p$ for the next iteration are updated as $\{1/(\sigma_j(\widehat{\mathbf{\Theta}}^{(\ell)}) + \varepsilon)\}_{j=1}^p$ so that a sequence of weights becomes in nondecreasing order. We repeat steps (II) and (III) until the rank of the estimated matrix does not change over the iteration ℓ.

Motivated from Candes, Wakin, and Boyd (2008), the rationale behind the weight updating rule (i.e., step (III)) is from the definition of weighted nuclear norm (3) with nondecreasing order (i.e., $0 \le w_1 \le \cdots \le w_p$); that is, the small weights are assigned for the large estimated singular values, whereas the large weights should be put on the small estimated singular values. The introduced parameter $\epsilon > 0$ in step (III) guarantees that, for any $j \in \{1, \ldots, p\}$, the $(\ell+1)$th updated weight $w_j^{(\ell+1)}$ is computable, when $\sigma_j(\widehat{\mathbf{\Theta}}^{(\ell)}) = 0$. Following Candes, Wakin, and Boyd (2008), we set $\varepsilon = 10^{-3}$, which works reasonably well in simulated settings in Section 5.

Remark 4. The weight update rule (I) $\sim$ (III) corresponds to solving (2) with a specific weight sequence $\boldsymbol{w} = \{1/(\sigma_1(\widehat{\mathbf{\Theta}}^{(\ell)}) + \varepsilon), \ldots, 1/(\sigma_p(\widehat{\mathbf{\Theta}}^{(\ell)}) + \varepsilon)\}$. Solving this problem is exactly equivalent to applying the Majorization-Minorization (MM) algorithm to the following surrogate objective function.

$$\min_{\mathbf{\Theta}\in\mathbb{R}^{d_1\times d_2}}\left\{\sum_{j=1}^{p}\log\left(\sigma_j(\mathbf{\Theta}) + \varepsilon\right)\right\} \qquad \text{s.t.} \qquad \boldsymbol{Y} = \boldsymbol{X}\mathbf{\Theta}.$$

Similarly noted by Candes, Wakin, and Boyd (2008), the log-sum penalty function $\sum_{j=1}^{p}\log\left(\sigma_j(\mathbf{\Theta}) + \varepsilon\right)$ has a potential to encourage more sparsity than the ℓ_1 norm on singular values (i.e., nuclear norm penalty) by allowing relatively large penalties to be placed on small singular values. This potentially implies solving (2) with weights updated via steps (I) $\sim$ (IV) can mitigate the rank over-estimation problem of SNN pointed out by Mukherjee

et al. (2015) and Bunea, She, and Wegkamp (2011). Readers can check this effect in the experiment presented in Section 5.2.

Remark 5. To the best of our knowledge, the ANN estimator from Chen, Dong, and Chan (2013) is the only paper that works on multi-variate linear regression problems via WNN. They suggest their way for the weight update rule. We compare the effects of the weight update rule in Section 4.1 with theirs on the estimation error. The detailed descriptions of the simulation setting and results are discussed in supplementary material I.

4.2. Surrogate Solution of $\widehat{\Theta}$ for the GCV Statistic

Let $\widehat{\boldsymbol{U}}\widehat{\boldsymbol{D}}\widehat{\boldsymbol{V}}^\top$ be the SVD of the converged solution from WMVR-ADMM in Section 2. Here, denote $\widehat{\mathbf{D}} := \operatorname{diag}(\widehat{d}_1, \ldots, \widehat{d}_p)$ as the diagonal matrix of singular values of the converged solution, which is not necessarily equal to the global minimizer $\widehat{\boldsymbol{\Theta}}$ of (2). Then, we define the following matrix $\boldsymbol{K} \in \mathbb{R}^{d_1 \times d_1}$: for any fixed weights $0 < w_1 < \cdots < w_{\widehat{r}}$, let $\boldsymbol{K} := \widehat{\boldsymbol{U}}_{\widehat{r}}\widehat{\boldsymbol{D}}^{\mathrm{K}}\widehat{\boldsymbol{U}}_{\widehat{r}}^\top := \sum_{j=1}^{\widehat{r}} \frac{w_j}{\widehat{d}_j}\widehat{\boldsymbol{U}}_j(\widehat{\boldsymbol{U}}_j)^\top$, where $\widehat{r}$ denotes the cardinality of a set $\{j : \widehat{d}_j > 0\}$ and $\widehat{\boldsymbol{U}}_j$ denotes the jth column of the matrix $\widehat{\boldsymbol{U}}$. With these notations, we provide the following proposition, whose proof is in Section F of supplemental material.

Proposition 4.1. For a fixed $\boldsymbol{K}$, denote $\widehat{\boldsymbol{\Theta}}^{\mathrm{SR}}$ as a minimizer of the following surrogate optimization problem:

$$\widehat{\boldsymbol{\Theta}}^{\mathrm{SR}} := \underset{\boldsymbol{\Theta}\in\mathbb{R}^{d_1\times d_2}}{\operatorname{argmin}} \left\{ \frac{1}{2n}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\Theta}\|_{\mathrm{F}}^2 + \frac{\lambda_n}{2}\operatorname{tr}\big(\boldsymbol{\Theta}^\top \boldsymbol{K}\boldsymbol{\Theta}\big) \right\}. \quad (13)$$

Then, under orthogonal design (i.e., $\boldsymbol{X}^\top\boldsymbol{X} = n\boldsymbol{I}_{d_1\times d_1}$), $\widehat{\boldsymbol{\Theta}}^{\mathrm{SR}} = \widehat{\boldsymbol{U}}^{\mathrm{LS}}\widehat{\boldsymbol{D}}^{\mathrm{SR}}\big(\widehat{\boldsymbol{V}}^{\mathrm{LS}}\big)^\top$, where $\widehat{\boldsymbol{D}}^{\mathrm{SR}}_{jj} = \widehat{d}_j$ for $j = 1, 2, \ldots, \widehat{r}$, and $\widehat{\boldsymbol{D}}^{\mathrm{SR}}_{jj} = \sigma_j\big(\widehat{\boldsymbol{\Theta}}^{\mathrm{LS}}\big)$ for $j = \widehat{r}+1, \ldots, p$.

Under the orthogonal design assumption, 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 $\widehat{\boldsymbol{\Theta}}^{\mathrm{LS}}$. Note that the assumption $\boldsymbol{X}^\top\boldsymbol{X} = n\boldsymbol{I}_{d_1\times d_1}$ requires the condition $n \geq d_1$. Although not theoretically justified, for some non-orthogonal random designs under the condition $d_1 > n$, we empirically observe the statement in Proposition 4.1 approximately holds. Specific simulation settings with results will be presented in supplementary material Section G. Note $\widehat{\boldsymbol{\Theta}}^{\mathrm{SR}}$ is not obtainable purely from data, but it plays a crucial role in estimating the degrees-of-freedom of $\widehat{\boldsymbol{\Theta}}$ in the next subsection.

4.3. GCV Statistic and Choice of Hyper-parameter λ_n

In this section, a GCV type of statistic (Golub, Heath, and Wahba 1979) for the choice of hyper-parameter λ_n is developed. Let us denote the $\widehat{\boldsymbol{\Theta}}^{\mathrm{W}}(\lambda_n)$ as a resulting estimator from the weight update procedure introduced in Section 4.1 with a fixed λ_n. Then, following Mukherjee et al. (2015), a GCV score for $\widehat{\boldsymbol{\Theta}}^{\mathrm{W}}(\lambda_n)$ from the multivariate linear regression problem is given by

$$\mathrm{GCV}(\lambda_n) := \frac{\operatorname{tr}\big(\big(\boldsymbol{Y} - \widehat{\boldsymbol{Y}}(\lambda_n)\big)\big(\boldsymbol{Y} - \widehat{\boldsymbol{Y}}(\lambda_n)\big)^\top\big)}{\big(nd_2 - \mathrm{df}(\widehat{\boldsymbol{Y}}(\lambda_n))\big)^2}, \quad (14)$$

where $\mathrm{df}(\widehat{\boldsymbol{Y}}(\lambda_n))$ is the degrees-of-freedom (Stein 1981; Efron 2004) of the model $\widehat{\boldsymbol{Y}}(\lambda_n) := \boldsymbol{X}\widehat{\boldsymbol{\Theta}}^{\mathrm{W}}(\boldsymbol{\lambda_n})$. The optimal $\lambda_n^\star$ is obtained when it minimizes the GCV score over the search range $\lambda_n \in [0, \mathcal{T}]$ for some $\mathcal{T} \geq 0$. The crux component in (14) is to compute the $\mathrm{df}(\widehat{\boldsymbol{Y}}(\lambda_n))$ either exactly or approximately. Motivated from Yuan et al. (2007) and the surrogate objective function (13), we can approximate $\widehat{\boldsymbol{Y}}(\lambda_n) := \boldsymbol{X}\widehat{\boldsymbol{\Theta}}^{\mathrm{W}}(\boldsymbol{\lambda_n})$ by

$$\widehat{\boldsymbol{Y}}(\lambda_n) \approx \boldsymbol{X}\Big(\boldsymbol{X}^\top\boldsymbol{X} + n\lambda_n\boldsymbol{K}\Big)^\dagger\boldsymbol{X}^\top\mathbf{Y},$$

where $(\cdot)^\dagger$ denotes the Moore-Penrose inverse of enclosed Gram-matrix. The degrees-of-freedom is approximated as follows:

$$\mathrm{df}(\widehat{\boldsymbol{Y}}(\lambda_n)) \approx d_2\operatorname{tr}\Big(\boldsymbol{X}\Big(\boldsymbol{X}^\top\boldsymbol{X} + n\lambda_n\boldsymbol{K}\Big)^\dagger\boldsymbol{X}^\top\Big), \quad (15)$$

where we use a simple fact: the degrees-of-freedom of a linear smoother, $\widehat{\boldsymbol{y}} := \boldsymbol{S}\boldsymbol{y}$ is $\mathrm{df}(\widehat{\boldsymbol{y}}) = \operatorname{tr}(\boldsymbol{S})$ for the smoothing matrix $\boldsymbol{S}$ and a vector $\boldsymbol{y}$. Note that the degrees-of-freedom is "approximated" in the sense that the non-linearities involved with $\boldsymbol{Y}$ in matrices $\boldsymbol{K}$ are ignored.

5. Numerical Experiments

Throughout this section, we generate the simulated data pair $(\boldsymbol{X}, \boldsymbol{Y})$ under the following setting. In Section 5.1, the coefficient matrix is generated from $\boldsymbol{\Theta}^\star = \boldsymbol{A}\boldsymbol{B}^\top \in \mathbb{R}^{d_1\times d_2}$, where each entry of $\boldsymbol{A} \in \mathbb{R}^{d_1\times r^\star}$ and $\boldsymbol{B} \in \mathbb{R}^{d_2\times r^\star}$ is from $\mathcal{N}(0,1)$. In Section 5.2, each entry of $\boldsymbol{\Theta}^\star \in \mathbb{R}^{d_1\times d_2}$ is independently sampled from $\mathcal{N}(0,1)$, and its first $r^\star$ singular values are replaced by values specified in the subsection, and rest are set as 0. For random design matrix $\boldsymbol{X} \in \mathbb{R}^{n\times d_1}$, each row of the matrix is sampled from $\mathcal{N}(0, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}_{i,j} = \xi^{|i-j|}$ for $i, j = 1, \ldots, d_1$. Here, the ξ is a parameter that controls correlations among features. The response matrix $\boldsymbol{Y}$ is generated from the model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\Theta}^\star + \boldsymbol{E}$, where the entries of $\boldsymbol{E}$ are independently from $\mathcal{N}(0,1)$. Thus, the simulation setting is characterized by the parameters $(n, r^\star, \xi, d_1, d_2)$.

5.1. Convergence of WMVR-ADMM

In this subsection, we present empirical evidence supporting the assertion made in Theorem 2.2; specifically, the simulations demonstrate that the solution obtained through WMVR-ADMM converges to a unique stationary point of Lagrangian function regardless of the initialized tuple $(\boldsymbol{\Theta}^{(0)}, \boldsymbol{\Gamma}^{(0)}, \boldsymbol{\Lambda}^{(0)})$. Additionally, we investigate whether the converged solution is a global minimum of the non-convex landscape (2) under the orthogonal design setting. The following two quantities are used to check the statements.

(I) For checking the Primal residual convergence (i.e., $\boldsymbol{\Theta}^{(k)} - \boldsymbol{\Gamma}^{(k)} \to 0$ as $k \to \infty$), and $\boldsymbol{\Gamma}^{(k)}$ convergence (i.e., $\boldsymbol{\Gamma}^{(k+1)} - \boldsymbol{\Gamma}^{(k)} \to 0$ as $k \to \infty$), we consider $R^{(k)} := \|\boldsymbol{\Theta}^{(k)} - \boldsymbol{\Gamma}^{(k)}\|_{\mathrm{F}}^2 + \|\boldsymbol{\Gamma}^{(k+1)} - \boldsymbol{\Gamma}^{(k)}\|_{\mathrm{F}}^2$.

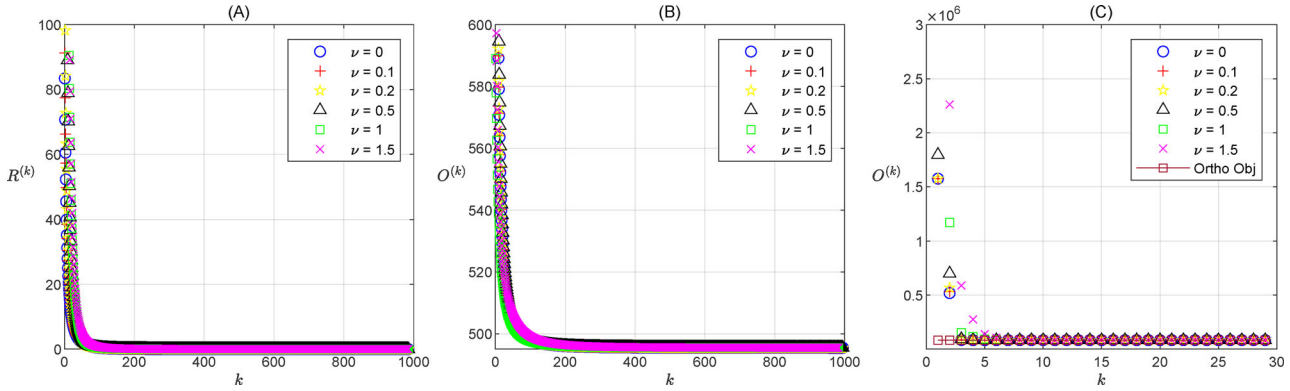

Figure 2. Convergences of $R^{(k)}$ (panel (A)) and $O^{(k)}$ (panel (B)) over the algorithm iteration index k. Panel (C) shows the empirical evidence for the global convergence of WMVR-ADMM under orthogonal design.

(II) For checking the objective convergence, we consider $O^{(k)} := \frac{1}{2n}\left\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\Theta}^{(k)}\right\|_{\mathrm{F}}^2 + \lambda_n \left\|\boldsymbol{\Theta}^{(k)}\right\|_{\boldsymbol{w},\star}$.

In simulation, a parameter tuple is set as $(n, r^\star, \xi, d_1, d_2) = (250, 50, 0, 250, 250)$. Note that with $\xi = 0$, then $\Sigma = \mathcal{I}_{d_1 \times d_1}$. The hyper-tuning parameter λ_n is fixed at $5\sqrt{\frac{d_1+d_2}{n}}$. We vary the initialized tuple $(\boldsymbol{\Theta}^{(0)}, \boldsymbol{\Gamma}^{(0)}, \boldsymbol{\Lambda}^{(0)})$ of WMVR-ADMM in Algorithm 1. The 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 $\{w_j\}_{j=1}^p$ are updated once through the rules in Section 4.1, and with the updated weights, $R^{(k)}$ and $O^{(k)}$ are recorded over $1 \le k \le 1000$, where k is an iteration index of the algorithm.

The first two panels, Figure 2(A) and (B), are the results under the design setting $\rho = 0$. As predicted by Theorem 2.2, the algorithm converges (i.e., Figure 2(A)), and the converged solution has the same objective values regardless of the initialization tuple (i.e., Figure 2(B)), which can be the evidence that the algorithm converges to a unique point of the Lagrangian function. In the last panel (i.e., Figure 2(C)), we work under the orthogonal design setting. The panel displays that the converged solutions from WMVR-ADMM have the same objective value as that from the global optimal solution of (2). This can be checked because the closed-form solution of the global minimizer of (2) is known from Proposition 3.1 under the orthogonal design setting. Note that the algorithm converges within 30 iterations much faster than it does under the non-orthogonal design.

5.2. Comparisons of Estimation Error with Other Methods

In this subsection, we compare the finite sample performance of the proposed method with several other popular approaches for multivariate linear regression. The methods that we compare include the following:

(a) WNN, the estimator obtained via WMVR-ADMM with weight update rule and GCV procedure introduced in Section 4;
(b) Naïve-WNN, the estimate is the same with the item (a) but with naïve GCV procedure;
(c) SNN, the estimate from the standard nuclear norm penalized least square method (Yuan et al. 2007) with GCV procedure in Section 4;
(d) Naïve-SNN, the estimate is the same with the item (c) but with the naïve GCV procedure;
(e) ANN, the estimate from the adaptive nuclear norm method (Chen, Dong, and Chan 2013) with weight update rule in Chen, Dong, and Chan (2013) and λ_n chosen from 10-fold cross-validation;
(f) RRRR, reduced ridge rank regression method (Mukherjee and Zhu 2011) with parameter λ_n and rank selected by 10-fold cross-validation.

The naïve GCV procedure in (b) and (d) chooses tuning parameter λ_n for which the GCV score in (14) is minimized, but with a naïve estimator of degrees-of-freedom for $\widehat{\boldsymbol{Y}}(\lambda_n)$; that is $\mathrm{df}(\widehat{Y}(\lambda_n)) = \widehat{r}(r_X + d_2 - \widehat{r})$, where r_X denotes the rank of design matrix $\boldsymbol{X}$. Here, $\widehat{r}(r_X + d_2 - \widehat{r})$ denotes a number of free parameters in a $d_1 \times d_2$ matrix with rank $\widehat{r}$ (Mukherjee et al. 2015; Bunea, She, and Wegkamp 2011). The ANN method in (e) is introduced in Section 1.1. The RRRR method in (f) is shown to be effective in estimating the singular value structure of $\boldsymbol{\Theta}^\star$ under high collinearity of $\boldsymbol{X}$. We provide the closed-form solutions of ANN and RRRR methods and descriptions on 10-fold cross-validation on data pair $(\boldsymbol{X}, \boldsymbol{Y})$ in Section H of supplementary material. Following Chen, Dong, and Chan (2013), the search range for choices of optimal λ_n is set as $[0, \sigma_1(\widehat{\boldsymbol{\Theta}}^{\mathrm{LS}})^3]$ where the range is splitted into 100 intervals with same length. For any methods listed in $(a) \sim (f)$, let $\widehat{\boldsymbol{\Theta}}^{(m)}$ be an optimal estimator of $\boldsymbol{\Theta}^\star$ obtained from an mth data pair $(\boldsymbol{X}^{(m)}, \boldsymbol{Y}^{(m)})$. Over 200 data pairs, $\mathcal{D} := \{(\boldsymbol{X}^{(m)}, \boldsymbol{Y}^{(m)})\}_{m=1}^{200}$, we record two quantities for measuring the performances of estimators listed in $(a) \sim (f)$: (I) $\|\widehat{\boldsymbol{\Theta}}^{(m)} - \boldsymbol{\Theta}^\star\|_F^2$ and (II) $\mathrm{Rank}(\widehat{\boldsymbol{\Theta}}^{(m)})$. Under these two metrics, we consider the following three models with $(n, d_1, d_2) = (20, 8, 8)$.

(a) For model I, each entry of $\boldsymbol{\Theta}^\star \in \mathbb{R}^{8\times 8}$ is independently sampled from $\mathcal{N}(0, 1)$, and its singular values are replaced by the values $(3, 2, 1, 0, 0, 0, 0, 0)$.
(b) Model II is the same as model I, but with the singular values $(5, 0, 0, 0, 0, 0, 0, 0)$.
(c) Model III is the same as model I, but with the singular values $(5, 5, 5, 5, 0, 0, 0, 0)$.

We analyze each model at three different correlation levels between predictors. The correlation parameters of $\boldsymbol{\Sigma}$ are set as $\xi = \{0, 0.5, 0.7\}$. For the given 9 scenarios, the results are

Table 1. A summary of mean and variance (in parentheses) of $\{||\widehat{\boldsymbol{\Theta}}^{(m)} - \boldsymbol{\Theta}^{\star}||_F^2\}_{m=1}^{200}$ and $\{\text{Rank}(\widehat{\boldsymbol{\Theta}}^{(m)})\}_{m=1}^{200}$ in the first and second row, respectively, over the methods $(a) \sim (f)$ over nine scenarios. Bold values represent the best result for each row (in terms of the mean values).

Scenarios		Results for the following methods					
		WNN	Naïve-WNN	SNN	Naïve-SNN	ANN	RRRR
I	$\xi = 0$	3.05 (1.00)	**2.97** (1.05)	3.14 (0.85)	3.36 (1.22)	3.27 (1.24)	4.53 (1.53)
		3.86 (0.77)	2.75 (0.70)	6.36 (0.72)	4.50 (1.72)	3.68 (0.66)	**3.13** (1.32)
	$\xi = 0.5$	3.85 (1.28)	**3.58** (1.14)	3.63 (1.01)	4.31 (1.55)	4.54 (1.74)	4.50 (1.51)
		3.47 (0.74)	2.31 (0.61)	5.91 (0.61)	**2.96** (1.57)	3.46 (0.62)	2.64 (1.09)
	$\xi = 0.7$	**4.99** (1.55)	5.07 (1.45)	5.15 (1.38)	6.34 (1.52)	6.53 (2.64)	6.31 (1.71)
		2.75 (0.80)	2.18 (0.46)	4.54 (1.17)	2.43 (0.78)	3.13 (0.67)	**2.88** (1.33)
II	$\xi = 0$	**1.15** (0.63)	1.16 (0.67)	2.30 (1.30)	3.60 (2.07)	1.31 (0.74)	1.37 (0.89)
		1.03 (0.16)	**1.00** (0.00)	2.87 (0.80)	1.45 (0.66)	1.31 (0.56)	1.07 (0.28)
	$\xi = 0.5$	**1.94** (1.20)	1.95 (1.20)	5.28 (2.58)	6.75 (3.68)	1.85 (1.30)	2.19 (1.76)
		1.00 (0.07)	**1.00** (0.00)	2.45 (0.86)	1.65 (0.72)	1.25 (0.53)	1.09 (0.41)
	$\xi = 0.7$	**2.22** (1.15)	**2.22** (1.15)	5.21 (1.92)	6.35 (2.45)	2.65 (2.16)	2.36 (1.45)
		1.00 (0.00)	**1.00** (0.00)	1.94 (0.67)	1.37 (0.50)	1.26 (0.47)	1.11 (0.51)
III	$\xi = 0$	3.95 (1.18)	3.96 (1.31)	**3.79** (0.99)	3.88 (1.05)	4.31 (1.65)	7.11 (1.99)
		5.10 (0.66)	**4.18** (0.55)	6.89 (0.61)	6.19 (1.20)	4.76 (0.59)	4.69 (1.11)
	$\xi = 0.5$	4.91 (1.42)	4.83 (1.50)	4.40 (1.20)	**4.33** (1.26)	6.32 (2.38)	6.28 (1.73)
		4.66 (0.68)	**4.05** (0.46)	6.38 (0.74)	5.20 (1.23)	4.59 (0.66)	4.59 (1.10)
	$\xi = 0.7$	8.22 (1.95)	8.77 (2.08)	**7.02** (1.65)	9.39 (2.56)	9.83 (3.55)	8.37 (1.70)
		3.56 (0.93)	2.95 (1.02)	5.29 (1.15)	2.99 (1.69)	**3.95** (0.64)	4.61 (1.62)

presented in Table 1. For each of the scenarios, we record the means (variances in parentheses) of $\{||\widehat{\boldsymbol{\Theta}}^{(m)} - \boldsymbol{\Theta}^{\star}||_F^2\}_{m=1}^{200}$ and $\{\text{Rank}(\widehat{\boldsymbol{\Theta}}^{(m)})\}_{m=1}^{200}$ in the first and second row, respectively, over the methods $(a) \sim (f)$. In what follows, we summarize the results with the insights gained from them. No methods $(a) \sim (f)$ dominate either in having the lowest estimation error or in estimating the true rank of $\boldsymbol{\Theta}^{\star}$ over the nine presented scenarios. However, WNN estimators with methods (a) and (b) give us the satisfying results on 11 out of 18 cases. Among the 7 cases where they did not show the best results, 3 cases come from the estimation error results in model III. Given the nature of WNN estimators with non-decreasing weights, this is not surprising, as model III considers the cases where all the singular values of $\boldsymbol{\Theta}^{\star}$ are equivalent.

Note that in model I where the singular values are in decreasing order, WNN estimators showed the best performances in estimation error. Regarding the rank estimation, SNN estimators from (c) and (d) tended to overestimate the rank of coefficient matrices (Bunea, She, and Wegkamp 2011; Mukherjee et al. 2015), whereas WNN estimators in (a) and (b) rather gave the underestimated ranks. Readers can revisit Section 4.1 where we provide a possible reason for this observation. In extreme cases where the rank of $\boldsymbol{\Theta}^{\star}$ is 1 (model II), WNN estimators both from (a) and (b) exhibited impressive performances where they almost accurately estimated the ranks over 200 estimators. Both ANN and RRRR estimators from (e) and (f) showed good performance in rank estimations across the nine scenarios (best or second to the best). In all three models, as the correlation parameter ξ increases, the estimation error increases.

5.3. Application to a Real Dataset

The proposed method is applied to an application, about a study of Polycyclic Aromatic Hydrocarbons (PAHs) from sec. 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.

The dataset includes 10 PAHs, which are pyrene (Py), acenaphthene (Ace), anthracene (Anth), acenaphthylene (Acy), chrysene (Chry), benzanthracene (Benz), fluoranthene (Fluora), fluorene (Fluore), naphthalene (Nap), and phenanthrene (Phen), and 25 complex 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 4 in Section L of supplemental material. From each of these mixtures, an electronic absorption spectrum is computed, The spectrum is digitized at 5 nm intervals in 27 wavelength channels from 220 nm to 350 nm, as shown in Panel (B) of Figure 4 in Section L of supplemental material. 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 sec. 5.1.2 of Brereton (2003) and sec. 2.2.2 of Isenmann (2008).

We are mainly interested in using WMVR-ADMM to understand the association between the concentrations from PAHs and the electronic absorption spectrum through the model (1). A figure demonstrates the concentrations from PAHs and the electronic absorption spectrum from the data are provided in supplemental material Section L. The method is conducted by following Algorithm 1, and the optimal tuning parameter λ_n and weights $\boldsymbol{w}$ are selected by the proposed GCV criterion described in Section 4. The resulting GCV scores are plotted in Figure 3(A) with respect to value λ_n, showing the selected λ_n is around 0.00021. The estimated eigenvalues with respect to λ_n are plotted in Figure 3(B), and under the optimal λ_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 3(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

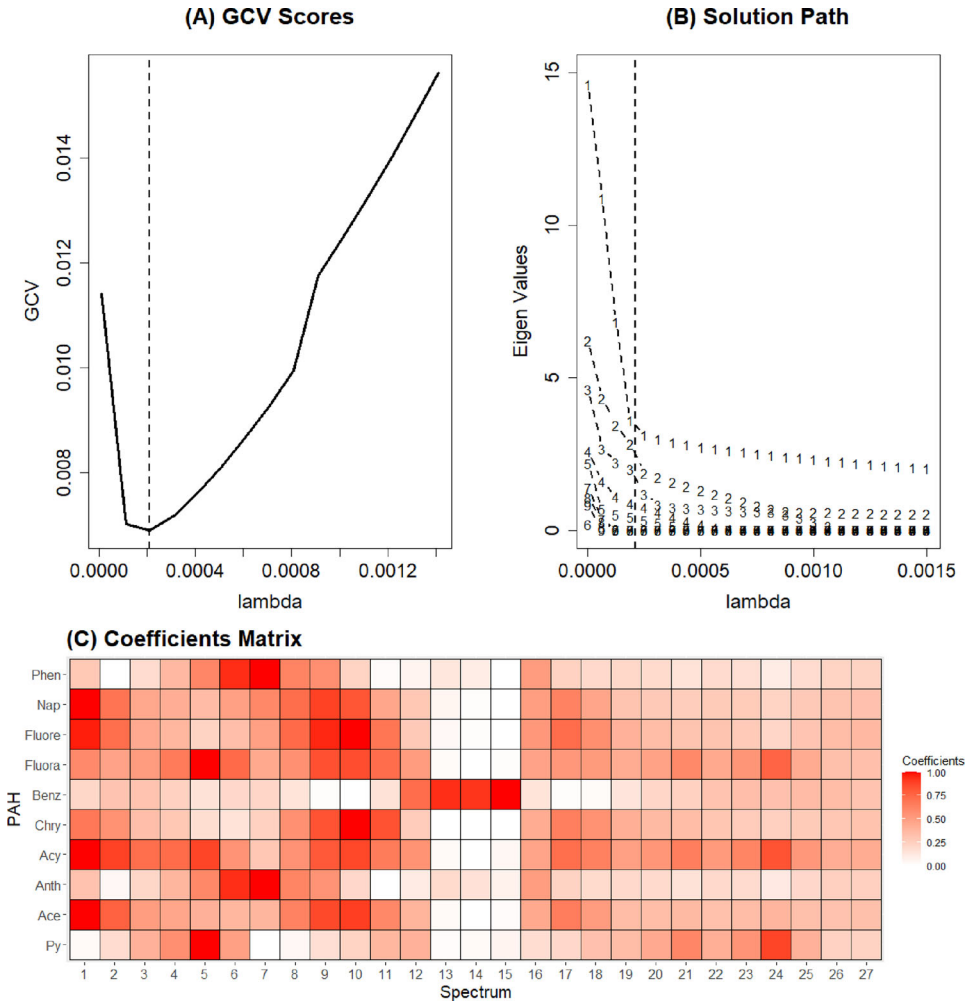

Figure 3. (A) GCV score versus tuning parameters λ, (B) solution path, (C) estimated coefficient matrix.

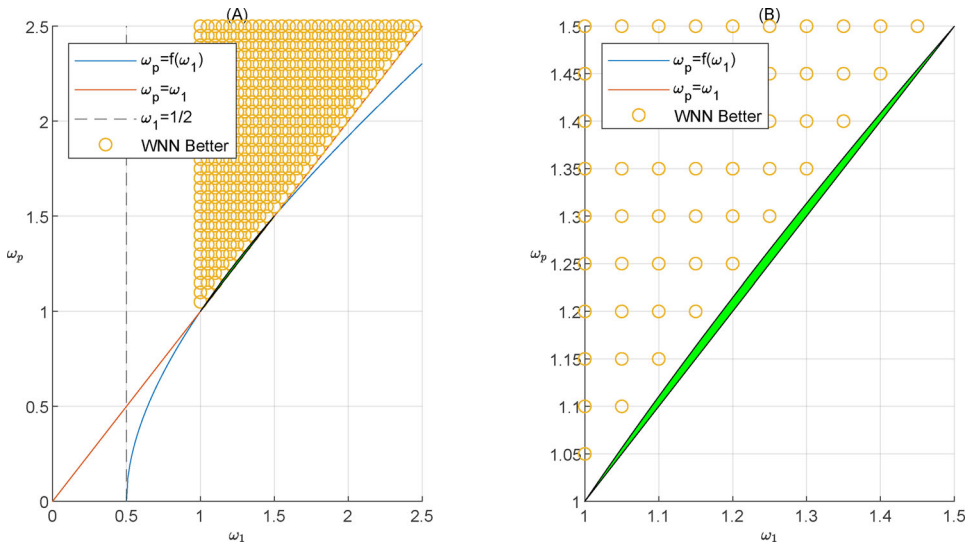

Figure 4. Panel (A) exhibits the intersected region of $\mathcal{W} \leq 3$ and $\frac{1}{2} < w_1 \leq \cdots \leq w_p$. Panel (B) magnifies the intersected region on grid $(w_1, w_p) \in [1, 1.5] \times [1, 1.5]$. Here, $f(w_1) := \frac{1}{4}(w_1 - \frac{1}{2}) + \frac{1}{4}\sqrt{w_1^2 + 23w_1 - \frac{47}{4}}$. Yellow circles represent the grid points where the WNN estimators give smaller estimation errors than SNN estimators do in our simulation setting.

relatively large. Additionally, these larger coefficients are usually from smaller column numbers in the heatmap. Thus, this shows the channels with smaller wavelengths are more important than larger wavelength channels.

6. Discussion

Several remaining open questions require further investigation 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 well-known 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 for future studies. Under the 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, Parikh, and Chu (2011). Nonetheless, we need further investigation whether these conditions can be used under our WNN setting with nondecreasing weights.
2. Although Theorem 3.3 demonstrates the estimation error from the WNN method converges to zero at the minimax rate, it's still unclear whether WNN outperforms SNN with a finite sample size. In Figure 4 the yellow circles represent the grid points of (w_1, w_p) over $[0, 2.5] \times [0, 2.5]$ where WNN gives smaller estimation errors than that by SNN with the tuning parameter λ_n and the sample size n being fixed. (More specific settings for simulations are deferred in supplementary material Section *K*.) The green colored region is an intersection of two constraints $\mathcal{W} \leq 3$ and $\frac{1}{2} < w_1 \leq w_p$. Here, the constraint $\mathcal{W} \leq 3$ is from Theorem 3.3 reflecting the set of pairs (w_1, w_p) for which WNN has a lower constant factor than SNN has. (Recall the definition of $\mathcal{W}$ and also recall SNN corresponds to the case $w_1 = w_p = 1$.) Note the green-colored region cannot cover the yellow circles which means that the result in Theorem 3.3 requires further refinements to justify our simulation result. We leave this as an open problem for future research.

Supplemental Materials

More technical details and numerical results are summarized in the supplemental material. The code for the numerical results and figures of the article and the associated user guidelines are also available in the supplemental material.

Acknowledgments

We deeply appreciate the insightful suggestions and comments from the editor Dr. Faming Liang, the associated editor, and the two reviewers to improve our article. We also want to express our appreciation to Prof. Bin Li from the Department of Experimental Statistics at Louisiana State University for his kind discussions and suggestions for our article writing.

Disclosure Statement

The authors report there are no competing interests to declare.

Funding

This material is based upon work supported by the National Science Foundation under grant no. 2229876 is partly supported by funds provided by the National Science Foundation, the Department of Homeland Security, and IBM. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or its federal agency and industry partners. The authors are also partially sponsored by NSF grants DMS 2015363 and the A. Russell Chandler III Professorship at Georgia Tech.

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. [4]

Bach, F. R. (2008), "Consistency of Trace Norm Minimization," *The Journal of Machine Learning Research*, 9, 1019–1048. [2]

Boyd, S., Parikh, N., and Chu, E. (2011), *Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers*, Delft: Now Publishers Inc. [2,11]

Brereton, R. G. (2003), *Chemometrics: Data Analysis for the Laboratory and Chemical Plant*, Chichester: Wiley. [9]

Bunea, F., She, Y., and Wegkamp, M. H. (2011), "Optimal Selection of Reduced Rank Estimators of High-Dimensional Matrices," *The Annals of Statistics*, 39, 1282–1309. [7,8,9]

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. [2,3,6]

Chen, K., Dong, H., and Chan, K.-S. (2013), "Reduced Rank Regression via Adaptive Nuclear Norm Penalization," *Biometrika*, 100, 901–920. [1,2,3,7,8]

Dong, W., Shi, G., Li, X., Ma, Y., and Huang, F. (2014), "Compressive Sensing via Nonlocal Low-Rank Regularization," *IEEE Transactions on Image Processing*, 23, 3618–3632. [2]

Efron, B. (2004), "The Estimation of Prediction Error: Covariance Penalties and Cross-Validation," *Journal of the American Statistical Association*, 99, 619–632. [7]

Fan, J., Gong, W., and Zhu, Z. (2019), "Generalized High-Dimensional Trace Regression via Nuclear Norm Regularization," *Journal of Econometrics*, 212, 177–202. [2]

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. [2]

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. [7]

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. [2]

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*, pp. 2862–2869. [2]

Isenmann, A. (2008), *Modern Multivariate Statistical Techniques: Regression, Classification and Manifold Learning*, New York: Springer. [9]

Kim, G., Cho, J., and Kang, M. (2020), "Cauchy Noise Removal by Weighted Nuclear Norm Minimization," *Journal of Scientific Computing*, 83, 1–21. [2,4]

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. [2,6]

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. [2]

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. [2]

Mukherjee, A., Chen, K., Wang, N., and Zhu, J. (2015), "On the Degrees of Freedom of Reduced-Rank Estimators in Multivariate Regression," *Biometrika*, 102, 457–477. [7,8,9]

Mukherjee, A., and Zhu, J. (2011), "Reduced Rank Ridge Regression and its Kernel Extensions," *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 4, 612–622. [3,8]

Negahban, S., and Wainwright, M. J. (2011), "Estimation of (near) Low-Rank Matrices with Noise and High-Dimensional Scaling," *The Annals of Statistics*, 39, 1069–1097. [2,4,6]

Rohde, A., and Tsybakov, A. B. (2011), "Estimation of High-Dimensional Low-Rank Matrices," *The Annals of Statistics*, 39, 887–930. [6]

Shang, P., and Kong, L. (2021), "Regularization Parameter Selection for the Low Rank Matrix Recovery," *Journal of Optimization Theory and Applications*, 189, 772–792. [11]

Stein, C. M. (1981), "Estimation of the Mean of a Multivariate Normal Distribution," *The Annals of Statistics*, 9, 1135–1151. [7]

Wang, Y., Yin, W., and Zeng, J. (2019), "Global Convergence of ADMM in Nonconvex Nonsmooth Optimization," *Journal of Scientific Computing*, 78, 29–63. [4]

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*, pp. 1096–1104. [2]

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*, pp. 3165–3174. [2]

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, 69, 329–346. [1,2,3,7,8]

Zha, Z., Yuan, X., Li, B., Zhang, X., Liu, X., Tang, L., and Liang, Y.-C. (2017), "Analyzing the Weighted Nuclear Norm Minimization and Nuclear Norm Minimization based on Group Sparse Representation," arXiv preprint arXiv:1702.04463. [2]

Zou, H. (2006), "The Adaptive Lasso and its Oracle Properties," *Journal of the American Statistical Association*, 101, 1418–1429. [3]