Electronic Journal of Statistics
Vol. 15 (2021) 4167–4191
ISSN: 1935-7524
https://doi.org/10.1214/21-EJS1895

Reduced rank regression with matrix projections for high-dimensional multivariate linear regression model

Wenxing Guo* **and Narayanaswamy Balakrishnan**

Department of Mathematics and Statistics, McMaster University, Hamilton, ON, L8S 4K1, Canada
e-mail: guow14@mcmaster.ca; bala@mcmaster.ca

Mengjie Bian

Department of Mathematics and Statistics, McMaster University, Hamilton, ON, L8S 4K1, Canada
e-mail: bianm1@mcmaster.ca

Abstract: In this work, we incorporate matrix projections into the reduced rank regression method, and then develop reduced rank regression estimators based on random projection and orthogonal projection in high-dimensional multivariate linear regression model. We propose a consistent estimator of the rank of the coefficient matrix and achieve prediction performance bounds for the proposed estimators based on mean squared errors. Finally, some simulation studies and a real data analysis are carried out to demonstrate that the proposed methods possess good stability, prediction performance and rank consistency compared to some other existing methods.

MSC2020 subject classifications: Primary 62F30, 62H12; secondary62J99.
Keywords and phrases: Matrix projection, reduced rank regression, dimension reduction, high-dimensional data, multivariate linear regression model.

Received June 2020.

Contents

*Corresponding author.

1. Introduction

Multivariate linear regression methods are widely used statistical tools in regression analysis. In general, a multivariate linear regression has n observations with r responses and p predictors, and can be expressed as

$$Y = XB + \varepsilon, \tag{1}$$

where $Y \in \mathbb{R}^{n\times r}$ denotes a multivariate response matrix, $X \in \mathbb{R}^{n\times p}$ represents a matrix of predictors, $\varepsilon \in \mathbb{R}^{n\times r}$ is an error matrix with its entry ε_{ij} being independent of each other with mean zero and variance σ_{ij}^2, and $B \in \mathbb{R}^{p\times r}$ is the regression coefficient matrix. The model in (1) is the foundation of multivariate regression analysis with its aim being to study the relationship between X and Y through the regression coefficient matrix B.

For model (1), the ordinary least-squares (OLS) estimator of B is

$$\hat{B}_{LS} = X^{+}Y, \tag{2}$$

where X^{+} denotes the Moore–Penrose inverse of X.

1.1. Existing work

The OLS method of multiple responses, under no constraints, is equivalent to performing OLS estimation for each response variable, separately, and so the estimator does not use the possible correlation between multiple responses. In practice, it will be quite realistic to assume that the response variables are correlated. One way of avoiding this drawback of the OLS method will be to consider reduced rank regression (RRR) model [19]. The reduced rank regression would allow the rank of B to be less than $min(p, r)$, and so the model parametrization can be expressed as $B = B_1B_2$, where $B_1 \in \mathbb{R}^{r\times d}$, $B_2 \in \mathbb{R}^{d\times p}$, and rank($B_1$)=rank($B_2$)=$d$. The decomposition $B = B_1B_2$ is non-unique since, for any orthogonal matrix $O \in \mathbb{R}^{d\times d}$, $B_1^* = B_1O$ and $B_2^* = O^TB_2$ will result in other valid decompositions satisfying $B = B_1^*B_2^* = B_1B_2$. Nevertheless, the parameter B of interest is identifiable, as well as span(B_1)=span(B) and

span(B_2^T)=span(B^T). Under some constraints on B_1 and B_2, such as $B_2 B_2^T = I_d$ or $B_1^T B_1 = I_d$, [2, 19, 21] derived the maximum likelihood estimators of the RRR parameters. As there are some linear constraints on regression coefficients, the number of effective parameters gets reduced and as a result the prediction accuracy may get improved. In high-dimensional data, a large number of predictor variables will be typically available, but some of them may not be useful for predictive purpose. Bunea et al. [4] proposed a rank selection criterion for selecting the optimal reduced rank estimator of the regression coefficient matrix in multivariate response regression models, and derived minimax optimal bounds based on mean squared errors of the estimators. Chen et al. [6] proposed an adaptive nuclear norm penalization method with low-rank matrix approximation, and developed a method for simultaneous dimension reduction and coefficient estimation in high-dimensional multivariate linear regression. If some column vectors of a predictor matrix X are nearly linearly dependent, the situation known as multicollinearity, the OLS estimator is known to perform poorly. Similarly, the performance of the reduced rank estimator is also not satisfactory when the predictor variables are highly correlated or the ratio of signal to noise is small. To overcome this problem, by incorporating ridge penalty into reduced rank regression, a reduced rank ridge regression estimator has been proposed [13, 6].

Dimension reduction is a way to reduce the number of random variables using various mathematical and statistical tools. In this regard, random projection is a widely used dimension reduction method in statistical and machine learning literature. Dobriban and Liu [8] examined different random projection methods in an unified framework, and derived explicit formulas for the accuracy loss of these methods compared to ordinary least-squares. Ahfock et al. [1] studied statistical properties of sketched regression algorithms and achieved new distributional results for a large class of sketched estimators and a conditional central limit theorem for the sketched dataset. Wang et al. [25] examined matrix ridge regression problems based on classical sketch and Hessian sketch. Thanei et al. [22] discussed some applications of random projections in linear regression models, and computational costs and theoretical guarantees of the generalization error in terms of random projection methods. Furthermore, random projection ideas have also been applied to problems in classification [5], clustering [9], and convex optimization [15, 16, 17] Principal components regression (PCR), based on principal components analysis, is also a classical tool for dimension reduction method, and so is the use of PCR to overcome the multicollinearity problem. Slawski [20] found a connection and made a comparison between principal components regression and random projection methods in classical linear regression.

1.2. Main contributions of this work

In this work, we propose three reduced rank estimators with a nuclear norm penalty in multivariate linear regression model in terms of single random projection, averaged random projection and principal components analysis, respectively. The estimation performance bounds of the proposed estimators are

achieved based on mean squared errors. Some simulation studies and a real data analysis are performed to demonstrate that the proposed estimators possess good stability and prediction performance compared to some other existing methods under certain conditions. In our model, the number of parameters p and r can be either less than the observed value n or greater than n. Moreover, the entry ε_{ij} in error matrix can have different variance σ_{ij}^2. Thus, the model considered here is a different one from those in Bunea et al. [4] and Chen et al. [6], in which the authors have assumed that all entries of the error matrix have the same variance σ^2. Thus, their models become a special case of the model considered here. We also develop a consistent estimation approach of the rank of the regression coefficient matrix, and the practical performance of the proposed rank estimation method is then demonstrated through simulation studies.

1.3. Notation

For a matrix $A \in \mathbb{R}^{n\times p}$, $\lambda_i(A)$ denotes the ith largest singular value of A. For $m,\ n \in \mathbb{R}$, $m \wedge n$ denotes $min\{m,\ n\}$. The Frobenious norm, nuclear norm and spectral norm of A are denoted by $||A||_F = \sqrt{tr(A^TA)}$, $||A||_* = \sum_{i=1}^{n\wedge p} \lambda_i(A)$ and $||A||_2 = \lambda_1(A)$, respectively. P_A denotes the orthogonal projection matrix $A(A^TA)^+A^T$. vec$(\cdot)$ operator transforms an $n \times m$ matrix into an nm-dimensional column vector by stacking the columns of the matrix below each other. $A \otimes B$ denotes the Kronecker product of two matrices A and B. Finally, $tr(\cdot)$ denotes the trace of a square matrix.

2. Reduced rank regression with matrix projections

Yuan et al. [27] proposed a reduced rank estimator with nuclear norm penalty by minimizing the penalized least squares criterion

$$\frac{1}{2} \parallel Y - XB \parallel_F^2 + \mathcal{P}_\mu(B), \tag{3}$$

where $\mathcal{P}_\mu(B) = \mu||B||_*$. The penalty produces sparsity among the singular values and thus achieves dimension reduction and shrinkage estimation simultaneously.

Chen et al. [6] developed a new method for simultaneous dimension reduction and coefficient estimation in high-dimensional multivariate regression in terms of an adaptive nuclear norm penalization. For this, they replaced the penalty function in (3) by $\mathcal{P}_\mu(B) = \mu||XB||_{*\omega}$, where $||XB||_{*\omega} = \sum_{i=1}^{n\wedge r} \omega_i\lambda_i(XB)$ denotes an adaptive nuclear norm of XB and ω_i's are non-negative weights.

2.1. Reduced rank regression with single random projection

The OLS estimator and the reduced rank estimator may both perform poorly when column vectors of a data matrix are highly correlated. To avoid this problem, in this work, we develop a two-step estimation method. First, a low-rank matrix is utilized to approximate the data matrix, and then a reduced rank regression is performed in terms of nuclear norm penalty. Motivated by the work of Halko et al. [10], we use the low-rank matrix approximation of X to be

$$\tilde{X} = QQ^T X, \tag{4}$$

where $XS = QR$, $S \in \mathbb{R}^{p\times k}$ is a standard Gaussian matrix which is a random matrix whose entries are independent standard normal variables, $Q \in \mathbb{R}^{n\times k}$ is a matrix with k orthonormal columns, and $R \in \mathbb{R}^{k\times k}$ is an upper triangular matrix with positive diagonal elements.

Inspired by Eq. (4) and the work of Chen et al. [6], a reduced rank estimator with nuclear norm penalty, based on random projection, can then be derived by minimizing the penalized least squares criterion

$$\frac{1}{2} \parallel Y - \tilde{X}B \parallel_F^2 + \mu||\tilde{X}B||_*. \tag{5}$$

The following proposition shows that a closed-form global minimizer of (5) can be found.

Proposition 2.1. *Let $\tilde{X}$ equal QQ^TX and $P_{\tilde{X}}Y$ have a singular value decomposition as $\tilde{U}\tilde{D}\tilde{V}^T$. Then, a global minimizer of (5) is given by*

$$\tilde{B} = \tilde{X}^+ Y \tilde{V} \tilde{D}^+ \tilde{D}_\mu \tilde{V}^T, \tag{6}$$

where $\tilde{D}_\mu$=diag$[\{\lambda_i(P_{\tilde{X}}Y) - \mu\}_+,\ i = 1, \cdots, n \wedge r]$, and diag$(\cdot)$ represents a diagonal matrix with the enclosed vector on its diagonal.

The result in Proposition 2.1 follows directly from Lemma 1. The rank of the coefficient matrix B, denoted by r_0, can be regarded as the number of effective linear combinations of predictor variables relating to response variables. In practice, we need to estimate the rank of B. (6) indicates that the quality of the rank estimator is related to the ratio of signal to noise and the value of k, and by combining the works of Bunea et al. [4] and Chen et al. [6], we develop here a method of rank estimation of B that can be expressed as

$$\tilde{r} = \max\Big\{i :\ \lambda_i(P_{\tilde{X}}Y) > \frac{k\mu}{\eta r_{\mathrm{x}}}\Big\}, \tag{7}$$

where k, r_{x} and η represent the number of columns of random projection matrix S, rank of predictor matrix X and a pre-specified constant, respectively. In practice, we can get the values of k and μ by cross-validation.

The following theorem shows that the rank selection method proposed recovers consistently the true rank r_0 under certain conditions.

Theorem 2.1. *Suppose the entries of* $\varepsilon \in \mathbb{R}^{n\times r}$ *are independent of each other and* $\varepsilon_{ij} \sim N(0, \sigma_{ij}^2)$. *Also, let* $\mu = \eta r_X(1+\theta)\sqrt{2V(P_{\tilde{X}}\varepsilon)\log(n+r)}/(k\delta)$ *and* $\lambda_{r_0}(\tilde{X}B) > \dfrac{2k\mu}{\eta r_X}$, *for any* $\theta > 0$. *Then, we have*

$$P(\tilde{r} \neq r_0) \longrightarrow 0 \text{ as } n + r \longrightarrow \infty.$$

Theorem 2.1 holds when n or r tends to infinity, and p is not restricted. Therefore, the rank consistency of the proposed estimator is effective for both classical and high-dimensional cases.

In order to evaluate the performance of the proposed estimators and demonstrate the results obtained, we need to decompose the predictor matrix X and construct two sub random matrices of S as follows.

For a matrix $X \in \mathbb{R}^{n\times p}$ with $\text{rank}(X) = q$ and $q \leq n \wedge p$, the singular value decomposition (SVD) can be expressed as

$$\begin{aligned} X &= \Gamma\Lambda P^T \\ &= (\Gamma_1, \quad \Gamma_2)\begin{pmatrix}\Lambda_1 & \\ & \Lambda_2\end{pmatrix}\begin{pmatrix}P_1^T \\ P_2^T\end{pmatrix}, \end{aligned} \tag{8}$$

where Γ is a $n \times q$ column orthonormal matrix, Γ_1 is the $n \times l$ matrix in which the columns are the top l left singular vectors of X, Γ_2 is similarly the $n\times(q-l)$ matrix in which the columns are the bottom $q-l$ left singular vectors of X, Λ is a $q \times q$ diagonal matrix, Λ_1 is the $l \times l$ diagonal matrix consisting of the top l singular values of X ($l < q$), Λ_2 is similarly the $(q-l)\times(q-l)$ matrix consisting of the bottom $q-l$ singular values of X, P is a $p\times q$ column orthonormal matrix, P_1 is the $p \times l$ matrix in which the columns are the top l right singular vectors of X, and P_2 is similarly the $p \times (q-l)$ matrix in which the columns are the bottom $q-l$ right singular vectors of X. Let $S_1 = P_1^T S$ and $S_2 = P_2^T S$. Then, S_1 and S_2 are $l \times k$ and $(q-l) \times k$ matrices, respectively. Further, S_1 and S_2 are independent since P_1 and P_2 are column orthonormal matrices.

In this section, we obtain results to bound the difference between the true value and its estimated value based on single random projection. Our analyses are separated into two parts. First, we describe bounds on the probability of a large deviation. Next, we present some information about expected values. When expectation is not taken, $\| XB - \tilde{X}\tilde{B} \|_F$ is indeed a random variable. In this situation, the following theorem gives error bound with a certain probability.

Theorem 2.2. *Suppose* $\tilde{X}$, $\tilde{B}$ *and* Λ_2 *are as defined in (4), (6) and (8), respectively. Further, let the entries* ε_{ij}*'s of the error matrix* ε *be independent of each other, each following normal distribution* $N(0, \sigma_{ij}^2)$, *and* $k \geq l + 4$, *with* l *being a non-negative integer. Then, for any* $p \times r$ *matrix* C *with* $r(C) \leq r_0$,

some $\delta \in (0,1]$ *and all* $\gamma \geq 1$, $t \geq 1$,

$$\| XB - \tilde{X}\tilde{B} \|_F \leq \| XB - \tilde{X}C \|_F + 2\mu(1+\delta)\sqrt{2r_0}$$

with failure probability at most $\exp\big(-\theta^2 \log(n+r)\big)$. *More specifically, setting* $C = B$, *we have*

$$\| XB - \tilde{X}\tilde{B} \|_F \leq \| \Lambda_2 \|_F \| B \|_F \sqrt{\frac{3lt^2(\gamma+1)^2}{k-l+1} + 1} + 2\mu(1+\delta)\sqrt{2r_0}$$

with failure probability at most $\exp\big(-\gamma^2/2\big) + \exp\big(-\theta^2 \log(n+r)\big) + t^{-(k-l)}$, *where* k *represents the number of columns of the random projection matrix* S.

The following theorem provides a bound for the expected error in the Frobenius norm.

Theorem 2.3. *Suppose* $\tilde{X}$ *and* $\tilde{B}$ *are as defined in (4) and (6), respectively. Then, for any* $p \times r$ *matrix* C *with* $r(C) \leq r_0$,

$$\begin{aligned} E[\| XB - \tilde{X}\tilde{B} \|_F] \leq\ & \Big[\| XB \|_F^2 - \frac{2k}{n} < XC,\ XB >_F + \frac{k}{n} \| XC \|_F^2 \Big]^{1/2} \\ & + 2\sqrt{2r_0} \Big\{ \Big[\frac{k}{n} \sum_{i=1}^{n} \sum_{j=1}^{r} \sigma_{ij}^2 \Big]^{1/2} + \mu \Big\}. \end{aligned} \tag{9}$$

More specifically, setting $C = B$, *we have*

$$\begin{aligned} E[\| XB - \tilde{X}\tilde{B} \|_F] \leq\ & \Big(1 - \frac{k}{n}\Big)^{1/2} \Big[\sum_{i=1}^{q} \lambda_i^2(X) \Big]^{1/2} \| B \|_F \\ & + 2\sqrt{2r_0} \Big\{ \Big[\frac{k}{n} \sum_{i=1}^{n} \sum_{j=1}^{r} \sigma_{ij}^2 \Big]^{1/2} + \mu \Big\}, \end{aligned} \tag{10}$$

where k *represents the number of columns of the random projection matrix* S.

The expected error bound in (10) reveals some interesting features. The bound depends on the value of k with the first term decreasing with increasing k and the second term increasing with increasing k. Thus, the choice of k balances the sum of the two terms, which can result in the value of sum being minimum.

To compare the expected error bounds derived by using the estimation of X (proposed methods) with not using the estimation of X for model (1), the following corollary is given.

Corollary 2.4. *Suppose* $\hat{B}_{NC}$ *is as defined in (26) and* $\Gamma = [\gamma_{ih}]_{n \times q}$. *Then,*

$$E[\| XB - X\hat{B}_{NC} \|_F] \leq 2\sqrt{2r_0} \Big\{ \Big[\sum_{i=1}^{n} \Big(\sum_{j=1}^{r} \sigma_{ij}^2 \Big) \Big(\sum_{h=1}^{q} \gamma_{ih}^2 \Big) \Big]^{1/2} + \mu \Big\}. \tag{11}$$

Remark 1. The value of the second term on the right hand side of (10) may be less than the value of the term on the right hand side of (11) and so the sum of the two terms on the right hand side of (10) may also be less than the value of the term on the right hand side of (11), especially when $\left[\sum_{i=1}^{q}\lambda_i^2(X)\right]^{1/2}$ or $\| B \|_F$ is small or σ_{ij}^2's are large.

2.2. Reduced rank regression with averaged random projection

We have studied reduced rank estimator with a single random projection. The variance of reduced rank estimator is an important problem in practical application. The variance of the estimator can be reduced by averaging multiple reduced rank estimators from different random projections. In this section, we propose a reduced rank regression with averaged random projections, inspired by the works of [3, 20, 22].

Definition 2.1. Let $\{S_m\}_{m=1}^{M} \in \mathbb{R}^{p\times k}$ be independent random projection matrices. Then, we define the averaged random projection matrix as

$$S^M = \frac{1}{M}\sum_{m=1}^{M} S_m. \tag{12}$$

Proposition 2.2. *Suppose* $\tilde{X}$, $\tilde{B}$ *and* $(\tilde{X}\tilde{B})^M$ *are as defined in (4), (6) and (12), respectively. We then have*

$$E[\| XB - (\tilde{X}\tilde{B})^M \|_F] \leq E[\| XB - \tilde{X}\tilde{B} \|_F]. \tag{13}$$

The proof is similar to the proof of Proposition 4 of Slawski [20]. The result in (13) suggests that reduced rank estimator with averaged random projection reduces the estimation error, improving significantly the efficiency of estimator.

2.3. Reduced rank regression with principal components analysis

For a data matrix X, let $\Gamma\Lambda P^T$ be the SVD of X as in (8). Then, the top k principal components X_k can be extracted from X, by setting $X_1 = XP_1$, where $P_1 \in \mathbb{R}^{p\times k}$ denotes the top k right singular vectors of X. We can then obtain a low-rank matrix approximation of X by using the top k right singular vectors of X as

$$\hat{X} = XP_1P_1^T. \tag{14}$$

Eq. (14) is different from Eq. (4). First, P_1 in (14) is a deterministic matrix derived by the principal components analysis of X, while Q in (4) is a random matrix obtained via QR factorization of X which is multiplied by a random matrix S from right. Second, $\hat{X}$ is achieved by multiplying X from right using the orthogonal projection matrix of P_1, and $\tilde{X}$ is obtained by orthogonally projecting X onto the column of Q.

By minimizing $\| Y - \hat{X}B \|_F^2$, we obtain the ordinary principal components regression estimator as

$$\hat{B}_{PC} = \hat{X}^+ Y = P_1 P_1^T \hat{B}_{LS}. \tag{15}$$

Further, a reduced rank estimator with nuclear norm penalty in terms of principal components analysis is obtained by minimizing the penalized least squares criterion

$$\frac{1}{2} \| Y - \hat{X}B \|_F^2 + \mu ||\hat{X}B||_*. \tag{16}$$

Proposition 2.3. *Let $\hat{X} = XP_1P_1^T$ and $P_{\hat{X}}Y$ have a singular value decomposition to be $\hat{U}\hat{D}\hat{V}^T$. Then, a minimizer of (16) is*

$$\hat{B} = \hat{X}^+ Y \hat{V} \hat{D}^+ \hat{D}_\mu \hat{V}^T, \tag{17}$$

where $\hat{D}_\mu = diag[\{\lambda_i(P_{\hat{X}}Y) - \mu\}_+,\ i = 1, \cdots, n \wedge r]$.
Similarly, the estimated rank of B can be expressed as

$$\hat{r} = \max\Big\{ i :\ \lambda_i(P_{\hat{X}}Y) > \frac{k\mu}{\eta r_X} \Big\}, \tag{18}$$

where k represents the number of principal components used.

Corollary 2.5. *Suppose $\hat{X}$ and $\hat{B}$ are as defined in (14) and (17), respectively. Then, for any $\theta > 0$ and some $\delta \in (0, 1]$,*

$$\| XB - \hat{X}\hat{B} \|_F \leq \Big[\sum_{i=k+1}^{q} \lambda_i^2(X) \Big]^{1/2} \| B \|_F + 2\mu(1+\delta)\sqrt{2r_0}$$

with failure probability at most $\exp\big(- \theta^2 \log(n + r)\big)$, where k represents the number of principal components used.

Corollary 2.6. *Suppose $\hat{X}$ and $\hat{B}$ are as defined in (14) and (17), respectively. Also, suppose $\Gamma_1 = [\gamma_{ih}]_{n \times k}$. Then,*

$$\begin{aligned} E[\| XB - \hat{X}\hat{B} \|_F] \ \leq\ & \Big[\sum_{i=k+1}^{q} \lambda_i^2(X) \Big]^{1/2} \| B \|_F \\ & + 2\sqrt{2r_0} \Big\{ \Big[\sum_{i=1}^{n} \Big(\sum_{j=1}^{r} \sigma_{ij}^2 \Big) \Big(\sum_{h=1}^{k} \gamma_{ih}^2 \Big) \Big]^{1/2} + \mu \Big\} \end{aligned} \tag{19}$$

where k represents the number of principal components used.

Remark 2. In practice, k should be less than or equal to q. When $k = q$, the right hand side of (19) is equal to the right hand side of (11). On the other hand, the value of the second term on the right hand side of (19) is less than the value of the term on the right hand side of (11) when $k < q$, and so the sum of the two terms on the right hand side of (19) may be less than the value of the term on the right hand side of (11).

3. Simulation study

In this section, we carry out a simulation study to compare the proposed methods with some known existing methods in terms of estimation accurary, prediction accuracy and performance of rank recovery. The simulated data are from the true model in (1). We consider two cases: One when both p and r are less than n, and another when p and r are greater than n. The following are the specific details of these two cases:

(a) The row vectors of the predictors X were independently generated from multivariate normal distribution $N(0, \Sigma_X)$, wherein the elements of Σ_X are composed of $\rho^{|i-j|}$, and $\rho^{|i-j|}$ denotes the correlation of the pairwise elements in the row vector of the predictor matrix. The coefficient matrix B is constructed as $B = cB_1B_2^T$, with $c > 0$, $B_1 \in \mathbb{R}^{p\times r_0}$ and $B_2 \in \mathbb{R}^{p\times r_0}$. Here, $c > 0$ is a prespecified constant, called signal intensity, in order to control the values of the entries of matrix B. All elements of B_1 and B_2 are derived from the uniform $(0, 1)$ distribution. The entry ε_{ij} of the error matrix ε are from $N(0, \sigma_{ij}^2)$, where σ_{ij} is derived from the uniform (a, b) distribution with $0 \le a < b$. We set the correlation coefficient ρ=0.1, 0.5 and 0.9, respectively, and the signal intensity c=0.5 and 0.05, respectively. In addition, we take the values of (a, b) as $(0, 1)$, $(2, 3)$ and $(4, 5)$, respectively. In this case, three scenarios are considered for performing the simulation study with Scenario 1: $\rho = 0.1, 0.5, 0.9$, $c = 0.5$ and $(a, b) = (0, 1)$; Scenario 2: $\rho = 0.1, 0.5, 0.9$, $c = 0.05$ and $(a, b) = (2, 3)$; Scenario 3: $\rho = 0.1, 0.5, 0.9$, $c = 0.05$ and $(a, b) = (4, 5)$. We then simulated 100 data sets consisting of $n = 60$, $p = r = 30$, $r_0 = 10$ in all these three scenarios;

(b) The setting is similar to that in (a), except that 100 data sets of simulation consist of $n = 30$, $p = r = 40$, $r_0 = 10$.

In the tables and figures presented, ANR, RRR and RAN denote adaptive nuclear norm penalized estimator [6], rank penalized estimator [4] and robustified adaptive nuclear norm penalized estimator [6], respectively. PNR, SNR and MSN represent nuclear norm penalized estimator with principal components analysis, single random projection and averaged random projection, respectively, while MRE represents the median rank estimate and correct rank recovery percentage.

We made use of cross-validation for selecting k and μ, and compared several different values of η based on different generated data. These results show that $\eta = 1/2$ is a good choice. Therefore, in the following comparisons, $\eta = 1/2$ is specified. ANR, RRR and RAN were computed by using the R package "rrpack". We implemented all proposed estimators in R, as well. For all the methods, the estimation accuracy of regression coefficient is measured by MSE(B)=$\| \hat{B} - B \|_F^2 /(pr)$. For the prediction accuracy of regression function, PM(XB)=$\| X\hat{B} - XB \|_F^2 /(nr)$ is used to measure ANR, RRR and RAN. Moreover, we utilize PM(XB)=$\| \hat{X}\hat{B} - XB \|_F^2 /(nr)$, PM(XB)=$\| \tilde{X}\tilde{B} - XB \|_F^2 /(nr)$ and PM(XB)=$\| (\tilde{X}\tilde{B})^M - XB \|_F^2 /(nr)$ to measure PNR, SNR and MSN, respectively.

Results for case (a): In terms of the estimation accuracy of regression coefficient, from Table 1 and Figure 1, we see that the estimation errors of all

methods grow with an increase in correlation coefficient ρ and all methods have similar performance when the ratio of signal to noise is large. The estimation errors of ANR and RRR methods still grow with an increase in correlation coefficient ρ when the ratio of signal to noise is small, while the estimation errors of PNR, SNR and MSN methods all decrease and, therefore, the proposed three methods perform better than other methods in this situation. Moreover, the estimation errors of ANR, RRR and RAN methods all increase as the ratio of signal to noise decreases for the three correlation coefficients used. The estimation errors for PNR, SNR and MSN methods also decrease as the ratio of signal to noise decreases when the correlation coefficient is large, but the estimation errors are smaller than other methods. Furthermore, we see that the performance of PNR, SNR and MSN methods is quite stable and similar to that of ANR, RRR and RAN methods with changes in correlation coefficient ρ. For the prediction accuracy of regression function, we see that all methods have similar performance when the ratio of signal to noise is large, but the prediction errors of all methods decrease with increasing values of correlation coefficient ρ. RRR method performs poorly especially when the ratio of signal to noise is small. In most situations, PNR and MSN methods perform better than other methods for all values of ρ when the ratio of signal to noise is small. The performance of SNR and MSN methods is good when there is a small correlation between the predictor variables, while PNR method performs well when the correlation coefficient ρ is large. For the performance of rank recovery, as seen in Table 1, the proposed methods are better than the existing ones in terms of median rank estimate and correct rank recovery percentage.

Results for case (b): In this high dimensional case, the estimation errors of all methods are relatively large and decrease with increasing values of correlation coefficient ρ when the signal strength is high based on estimation accuracy; yet, the proposed methods have smaller estimation errors compared to other methods in this situation. Other comparisons are similar to those for case (a) in terms of estimation accurary, prediction accuracy and the performance of rank recovery.

4. Illustrative example

A breast cancer dataset was first used by Chin et al. [7]. It contains 89 samples comprising gene expression measurements and comparative genomic hybridization measurements. This dataset has been analyzed by Witten et al. [26] and Chen et al. [6], and these data are available in the R package PMA. It has been shown that some types of cancer have the characteristics of abnormal alterations of DNA copy number [14]. It will, therefore, be of interest to identify the relationship between DNA copy numbers and RNA expression levels. Here, we regress copy-number variations on gene expression profile since the prediction model can identify copy-number changes related to function. In this case, we consider chromosome 18, where $p = 294$, $r = 51$ and $n = 89$. We centered and scaled both predictor matrix X and response matrix Y. For comparison of

TABLE 1

Comparisons of different methods based on 100 simulation runs with n=60, p=30, r=30

		ANR	RRR	RAN	PNR	SNR	MSN
				S_1			
$\rho = 0.1$	MSE (B)	0.005 (0.000)	0.006 (0.000)	0.005 (0.000)	0.005 (0.000)	0.005 (0.000)	0.005 (0.000)
	PM (XB)	0.090 (0.000)	0.107 (0.000)	0.090 (0.000)	0.092 (0.000)	0.092 (0.000)	0.092 (0.000)
	MRE (B)	10, 38%	8, 10%	10, 41%	10, 55%	10, 57%	10, 62%
$\rho = 0.5$	MSE (B)	0.007 (0.000)	0.009 (0.000)	0.007 (0.000)	0.007 (0.000)	0.007 (0.000)	0.007 (0.000)
	PM (XB)	0.087 (0.000)	0.105 (0.000)	0.087 (0.000)	0.090 (0.000)	0.090 (0.000)	0.090 (0.000)
	MRE (B)	9, 32%	8, 5%	9, 36%	10, 66%	10, 63%	10, 68%
$\rho = 0.9$	MSE (B)	0.023 (0.000)	0.029 (0.000)	0.023 (0.000)	0.023 (0.000)	0.020 (0.000)	0.015 (0.000)
	PM (XB)	0.067 (0.000)	0.084 (0.000)	0.067 (0.000)	0.072 (0.000)	0.074 (0.000)	0.071 (0.000)
	MRE (B)	6, 1%	4, 0%	7, 2%	6, 0%	7, 0%	7, 0%
				S_2			
$\rho = 0.1$	MSE (B)	0.010 (0.000)	0.015 (0.000)	0.009 (0.000)	0.009 (0.000)	0.009 (0.000)	0.008 (0.000)
	PM (XB)	0.229 (0.002)	0.236 (0.002)	0.229 (0.002)	0.369 (0.003)	0.369 (0.003)	0.259 (0.003)
	MRE (B)	2, 0%	1, 0%	2, 0%	10, 55%	10, 51%	10, 61%
$\rho = 0.5$	MSE (B)	0.015 (0.000)	0.017 (0.000)	0.009 (0.000)	0.005 (0.000)	0.006 (0.000)	0.004 (0.000)
	PM (XB)	0.225 (0.002)	0.227 (0.002)	0.223 (0.002)	0.340 (0.003)	0.428 (0.004)	0.293 (0.003)
	MRE (B)	1, 0%	1, 0%	1, 0%	9, 21%	9, 28%	9, 33%
$\rho = 0.9$	MSE (B)	0.077 (0.013)	0.092 (0.033)	0.015 (0.000)	0.001 (0.000)	0.002 (0.000)	0.001 (0.000)
	PM (XB)	0.223 (0.002)	0.216 (0.002)	0.222 (0.002)	0.227 (0.002)	0.422 (0.003)	0.253 (0.002)
	MRE (B)	1, 0%	1, 0%	1, 0%	3, 0%	4, 0%	4, 0%
				S_3			
$\rho = 0.1$	MSE (B)	0.020 (0.000)	0.080 (0.002)	0.016 (0.000)	0.014 (0.000)	0.015 (0.000)	0.013 (0.000)
	PM (XB)	0.534 (0.015)	1.344 (0.168)	0.521 (0.012)	0.485 (0.009)	0.505 (0.009)	0.453 (0.006)
	MRE (B)	1, 0%	1, 0%	1, 0%	11, 16%	11, 10%	10, 31%
$\rho = 0.5$	MSE (B)	0.034 (0.001)	0.083 (0.002)	0.016 (0.000)	0.008 (0.000)	0.011 (0.000)	0.008 (0.000)
	PM (XB)	0.656 (0.021)	0.944 (0.038)	0.654 (0.021)	0.532 (0.010)	0.722 (0.020)	0.651 (0.012)
	MRE (B)	2, 0%	1, 0%	2, 0%	10, 100%	10, 100%	10, 100%
$\rho = 0.9$	MSE (B)	0.240 (0.021)	0.259 (0.015)	0.016 (0.000)	0.002 (0.000)	0.004 (0.000)	0.002 (0.000)
	PM (XB)	0.710 (0.028)	0.725 (0.017)	0.708 (0.028)	0.511 (0.013)	0.836 (0.027)	0.672 (0.019)
	MRE (B)	1, 0%	1, 0%	1, 0%	3, 0%	3, 0%	3, 0%

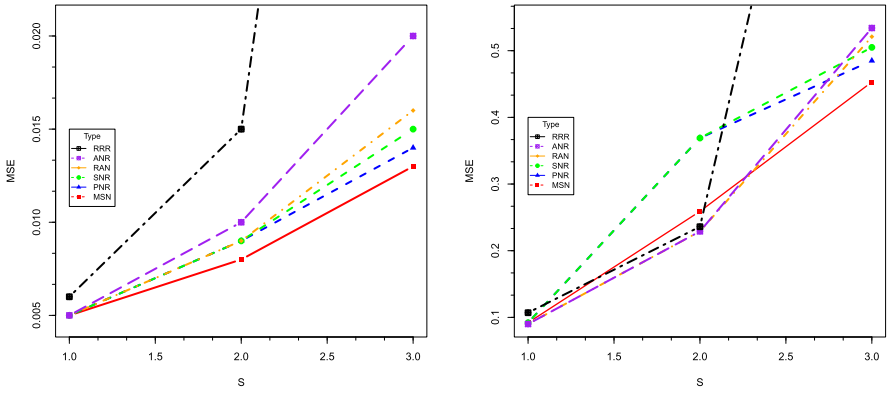
S_1, S_2 and S_3 denote Scenario 1, Scenario 2 and Scenario 3, respectively. The numbers in parentheses are the corresponding standard deviations.

TABLE 2

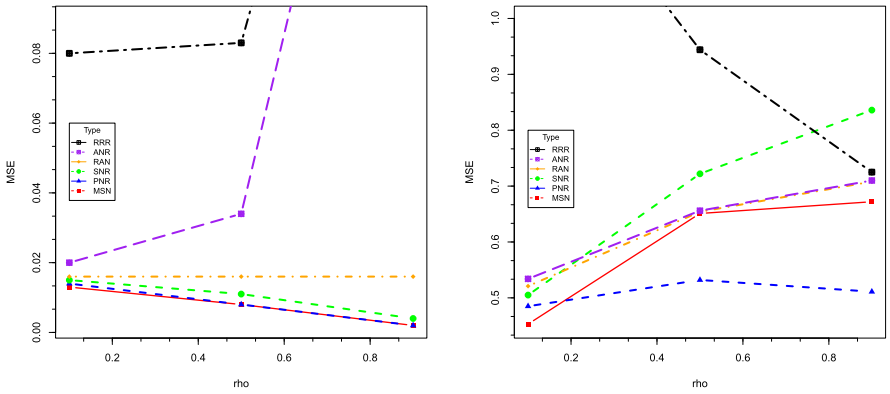
Comparisons of different methods based on 100 simulation runs with n=30, p=40, r=40.

		ANR	RRR	RAN	PNR	SNR	MSN
				S_1			
$\rho = 0.1$	MSE (B)	0.413 (0.023)	0.415 (0.023)	0.413 (0.022)	0.397 (0.022)	0.397 (0.022)	0.395 (0.021)
	PM (XB)	0.152 (0.000)	0.196 (0.001)	0.152 (0.000)	0.153 (0.000)	0.153 (0.000)	0.151 (0.000)
	MRE (B)	9, 22%	6, 2%	9, 22%	10, 60%	10, 57%	10, 62%
$\rho = 0.5$	MSE (B)	0.159 (0.004)	0.161 (0.004)	0.159 (0.004)	0.143 (0.003)	0.143 (0.003)	0.141 (0.003)
	PM (XB)	0.149 (0.000)	0.192 (0.001)	0.149 (0.000)	0.147 (0.000)	0.147 (0.000)	0.145 (0.000)
	MRE (B)	8, 17%	6, 2%	8, 20%	10, 57%	10, 65%	10, 69%
$\rho = 0.9$	MSE (B)	0.052 (0.000)	0.059 (0.001)	0.051 (0.000)	0.043 (0.000)	0.040 (0.000)	0.036 (0.000)
	PM (XB)	0.114 (0.000)	0.146 (0.001)	0.114 (0.000)	0.120 (0.000)	0.130 (0.000)	0.120 (0.000)
	MRE (B)	6, 2%	4, 0%	6, 3%	11, 33%	10, 59%	10, 63%
				S_2			
$\rho = 0.1$	MSE (B)	0.017 (0.000)	0.032 (0.002)	0.013 (0.000)	0.012 (0.000)	0.012 (0.000)	0.011 (0.000)
	PM (XB)	0.346 (0.008)	0.589 (0.557)	0.345 (0.005)	0.390 (0.005)	0.427 (0.007)	0.397 (0.004)
	MRE (B)	2, 0%	1, 0%	2, 0%	10, 51%	10, 53%	10, 58%
$\rho = 0.5$	MSE (B)	0.024 (0.000)	0.046 (0.011)	0.014 (0.000)	0.007 (0.000)	0.008 (0.000)	0.006 (0.000)
	PM (XB)	0.366 (0.007)	0.472 (0.327)	0.364 (0.008)	0.420 (0.005)	0.523 (0.009)	0.480 (0.006)
	MRE (B)	2, 0%	2, 0%	1, 0%	11, 5%	12, 2%	11, 6%
$\rho = 0.9$	MSE (B)	0.101 (0.005)	0.126 (0.115)	0.016 (0.000)	0.002 (0.000)	0.003 (0.000)	0.002 (0.000)
	PM (XB)	0.397 (0.008)	0.480 (0.507)	0.395 (0.004)	0.347 (0.005)	0.569 (0.007)	0.470 (0.005)
	MRE (B)	1, 0%	1, 0%	2, 0%	4, 0%	5, 0%	5, 0%
				S_3			
$\rho = 0.1$	MSE (B)	0.024 (0.000)	0.148 (0.012)	0.017 (0.000)	0.015 (0.000)	0.016 (0.000)	0.014 (0.000)
	PM (XB)	0.773 (0.037)	2.671 (4.798)	0.723 (0.028)	0.676 (0.022)	0.685 (0.038)	0.641 (0.021)
	MRE (B)	1, 0%	1, 0%	1, 0%	10, 100%	10, 100%	10, 100%
$\rho = 0.5$	MSE (B)	0.043 (0.001)	0.140 (0.036)	0.016 (0.000)	0.011 (0.000)	0.013 (0.000)	0.010 (0.000)
	PM (XB)	1.037 (0.059)	1.903 (3.380)	1.020 (0.043)	0.917 (0.028)	1.076 (0.035)	1.007 (0.034)
	MRE (B)	2, 0%	1, 0%	2, 0%	8, 0%	12, 3%	11, 5%
$\rho = 0.9$	MSE (B)	0.298 (0.047)	0.336 (0.062)	0.016 (0.000)	0.003 (0.000)	0.005 (0.000)	0.003 (0.000)
	PM (XB)	1.212 (0.052)	1.285 (0.053)	1.210 (0.050)	1.010 (0.035)	1.368 (0.080)	1.184 (0.046)
	MRE (B)	1, 0%	1, 0%	1, 0%	3, 0%	4, 0%	4, 0%

S_1, S_2 and S_3 denote Scenario 1, Scenario 2 and Scenario 3, respectively. The numbers in parentheses are the corresponding standard deviations.



(a) B estimators with S_1, S_2 , S_3 and $\rho = 0.1$ (b) XB estimators with S_1, S_2 , S_3 and $\rho = 0.1$



(c) B estimators with ρ=0.1, 0.5, 0.9 and S_3 (d) XB estimators with ρ=0.1, 0.5, 0.9 and S_3

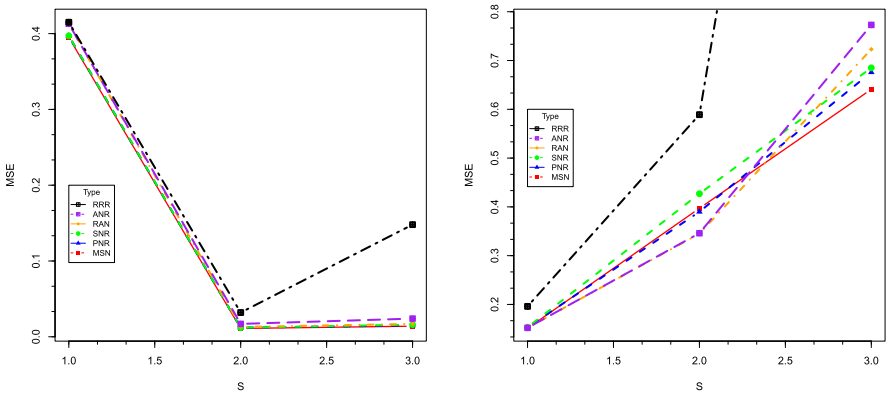
Fig 1: Comparisons of MSE of B and XB estimators based on 100 simulation runs with n=60, p=30, r=30.

prediction accuracy, a prediction mean squared error (PMSE) is defined as

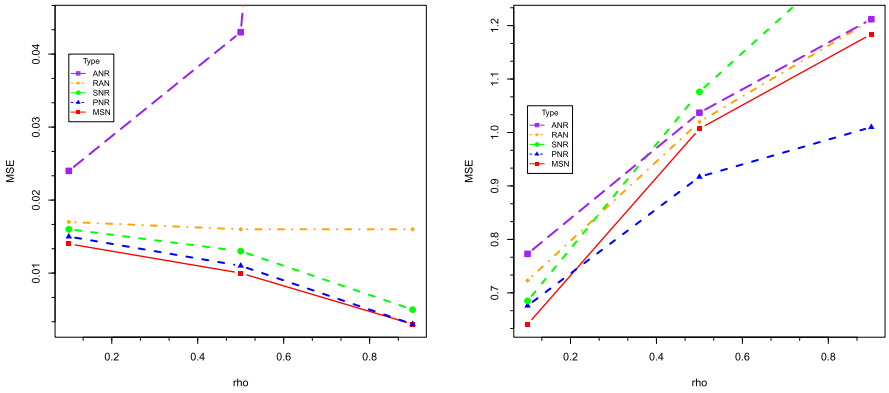
$$\text{PMSE} = \| Y_t - X_t \hat{B} \|_F^2 / (n_t r), \tag{20}$$

where (Y_t, X_t) represents the test dataset and $\hat{B}$ represents the estimator of B corresponding to each method. In addition, we randomly split the data into a training set of size 70 and a test set of size 19. The training dataset is used to achieve the estimation in the model, and then the test dataset is used to evaluate the prediction performance of estimators. All the tuning parameters were selected by ten-fold cross-validation.

As seen in Table 3 and Figure 3, the proposed estimators PNR, SNR and MSN are better than other estimators in terms of prediction performance and stability.



(a) B estimators with S_1, S_2 , S_3 and $\rho = 0.1$ (b) XB estimators with S_1, S_2 , S_3 and $\rho = 0.1$



(c) B estimators with ρ=0.1, 0.5, 0.9 and S_3 (d) XB estimators with ρ=0.1, 0.5, 0.9 and S_3

Fig 2: Comparisons of MSE of B and XB estimators based on 100 simulation runs with n=30, p=40, r=40.

More specifically, the prediction error of MSN method is the smallest and is the most stable one, followed by PNR method, while RRR method performs poorly in this case. Although the RAN method is better than the ANR method, it is not as good as the methods proposed in this work.

5. Discussion

We are considering the model exactly as considered in Bunea et al. [4] and Chen et al. [6], wherein the rows assume independence between elements. However, we allow for heterogeneity among the components in terms of different variances, which generalizes their model. It will be of interest to consider dependence between components within the rows, and this is something we wish to consider

TABLE 3
Prediction comparisons based on data split at random 100 times

	ANR	RRR	RAN	PNR	SNR	MSN
Rank	21	14	21	26	26	25
	(6.7)	(22.1)	(5.8)	(0.0)	(0.0)	(0.0)
PMSE	0.782	0.798	0.693	0.602	0.620	0.582
	(0.013)	(0.012)	(0.012)	(0.014)	(0.015)	(0.008)

Rank and PMSE represent the estimated rank and the prediction mean squared error, respectively. The numbers in parentheses are the corresponding standard deviations.

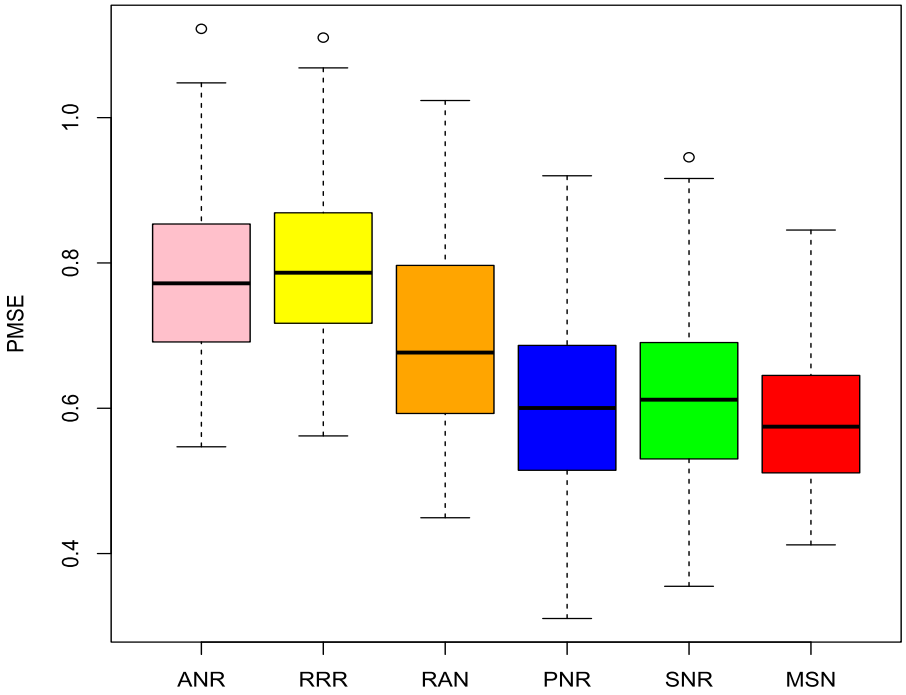


Fig 3: The distribution of PMSE based on data split at random 100 times

for our future work.

A weighted regression scheme can be taken into account in the following manner. Let $\Sigma = [\sigma_{ij}]_{n\times r}$ and Σ^j=[diag(σ_j)]$^{-1}$, where diag$(\cdot)$ represents a diagonal matrix with the enclosed vector on its diagonal, and σ_j denotes the jth column vector of Σ, $j = 1, \cdots, r$. Let I^j be an $r \times r$ matrix with the (jj)th entry being 1 and remaining entries being 0.

Consider the model

$$\sum_{j=1}^{r} \Sigma^j Y I^j = \sum_{j=1}^{r} \Sigma^j X B I^j + \sum_{j=1}^{r} \Sigma^j \varepsilon I^j. \tag{21}$$

Eq. (21) can be expressed as

$$Y^* = X^* B^* + \varepsilon^*, \tag{22}$$

where $Y^* = (\Sigma^1 Y, \cdots, \Sigma^r Y)(I^1, \cdots, I^r)^T$, $X^* = (\Sigma^1 X, \cdots, \Sigma^r X)$, $B^* = (I^1 B^T, \cdots, I^r B^T)^T$, and $\varepsilon^* = (\Sigma^1 \varepsilon, \cdots, \Sigma^r \varepsilon)(I^1, \cdots, I^r)^T$.

Using the transformation, we have that $\varepsilon^* \in \mathbb{R}^{n\times r}$ is an error matrix with its entries ε^*_{ij}'s being independent of each other with mean zero and variance 1. In particular, when $r = 1$, Eq. (22) reduces to the equation

$$\Sigma^1 Y = \Sigma^1 X B + \Sigma^1 \varepsilon. \tag{23}$$

This is an ordinary weighted linear regression model, where Y and ε reduce to $n \times 1$ vectors and B reduces to a $p \times 1$ vector.

Here, we refer to Eq. (22) as a multivariate weighted linear regression model compared to Eq. (23). Using model (22), weighted counterparts of proposed estimators can be obtained. Moreover, suppose σ_{ij}'s have the common factor σ_0, with $\sigma_{ij} = \sigma_0 \sigma^0_{ij}$ and Σ^j=[diag$(\sigma^0_j)]^{-1}$, and $\sigma^0_j = \sigma_0^{-1}\sigma_j$. We then have $\varepsilon^* \in \mathbb{R}^{n\times r}$ as an error matrix with its entries ε^*_{ij}'s being independent of each other with mean zero and variance σ_0^2, which is similar to the error matrix in Bunea et al. [4] and Chen et al. [6].

Now notice that the model (1) may be expressed in the vector form as

$$\mathrm{vec}(Y) = (I_r \otimes X)\,\mathrm{vec}(B) + \mathrm{vec}(\varepsilon), \tag{24}$$

where vec(Y) and vec(ε) are $nr\times 1$ vectors, B is a $pr\times 1$ vector, and Cov(vec(ε))= [diag(vec(Σ))]^2. Thus, model in (24) is a general linear regression model with heteroscedasticity.

Also, the model in (21) and (22) may be expressed in the vector form as

$$W\,\mathrm{vec}(Y) = W(I_r \otimes X)\,\mathrm{vec}(B) + W\,\mathrm{vec}(\varepsilon), \tag{25}$$

where $W = \sum_{j=1}^{r}(I^j \otimes \Sigma^j)$ is a weighted matrix and $\mathrm{Cov}\big(W\,\mathrm{vec}(\varepsilon)\big)=I_r \otimes I_n$. If σ_{ij}'s have the common factor σ_0, then $\mathrm{Cov}\big(W\,\mathrm{vec}(\varepsilon)\big)=\sigma_0^2 I_r \otimes I_n$.

In practice, when σ_{ij}'s are unknown, we can utilize some existing methods (such as [18, 11]) to estimate σ_{ij}'s in terms of model in (24).

Appendix A: Some useful lemmas

In order to achieve the results of propositions, theorems and corollaries, we introduce the following lemmas.

Lemma 1. *Let $P_X Y$ have a singular value decomposition, $P_X Y = UDV^T$, and for any $\mu \geq 0$, a global optimal solution of the representation*

$$\min_B \left\{ \frac{1}{2} \parallel Y - XB \parallel_F^2 + \mu ||XB||_* \right\} \tag{26}$$

is $X\hat{B}_{NC} = UD_\mu V^T$ *or* $\hat{B}_{NC} = \hat{B}_{LS}VD^+D_\mu V^T$, *where* $D_\mu = diag[\{\lambda_i(P_XY) - \mu\}_+,\ i = 1, \cdots, n \wedge r]$.

Proof. From the fact that $\| Y - XB \|_F^2 = \| Y - P_XY \|_F^2 + \| P_XY - XB \|_F^2$, we have (26) to be equivalent to

$$\min_B \Big\{ \frac{1}{2} \| P_XY - XB \|_F^2 + \mu ||XB||_* \Big\}. \tag{27}$$

In addition, we have $\| P_XY - XB \|_F^2 = tr(Y^TP_XY) - 2tr(P_XYB^TX^T) + tr(B^TX^TXB)$. Using von Neumann's trace inequality [24, 12], we obtain

$$tr(P_XYB^TX^T) \le \sum_{i=1}^{n\wedge r} \lambda_i(P_XY)\lambda_i(XB),$$

with the equality holding when XB satisfies the singular value decomposition $XB = Udiag[\lambda_i(XB)]V^T$. Thus, (27) can be re-expressed as

$$\min_B \Big\{ \sum_{i=1}^{n\wedge r} \Big[\frac{1}{2}\lambda_i^2(XB) - \big\{\lambda_i(P_XY) - \mu\big\}\lambda_i(XB) + \frac{1}{2}\lambda_i^2(P_XY) \Big] \Big\}. \tag{28}$$

It is obvious that the objective function in (28) can be minimized if $\lambda_i(XB) = \{\lambda_i(P_XY) - \mu\}_+$. Because $\{\lambda_i(P_XY)\}$ is non-increasing sequence, $X\hat{B}_{NC} = UD_\mu V^T$ is a global optimal solution. □

Lemma 2. *Assume that there exists an index* $m \le r_0$ *such that* $\lambda_{m+1}(\tilde{X}B) \le \dfrac{(1-\delta)k\mu}{\eta r_x}$ *and* $\lambda_m(\tilde{X}B) > \dfrac{(1+\delta)k\mu}{\eta r_x}$ *for some* $\delta \in (0, 1]$. *Then,*

$$P(\tilde{r} \ne m) \le P\Big\{ \lambda_1(P_{\tilde{X}}\varepsilon) \ge \frac{\delta k\mu}{\eta r_x} \Big\}.$$

Proof. From (7), we have $\tilde{r} > m \iff \lambda_{m+1}(P_{\tilde{X}}Y) > \dfrac{k\mu}{\eta r_{\mathrm{x}}}$ and $\tilde{r} < m \iff \lambda_m(P_{\tilde{X}}Y) \le \dfrac{k\mu}{\eta r_{\mathrm{x}}}$. It implies that

$$P(\tilde{r} \ne m) = P\Big\{ \lambda_m(P_{\tilde{X}}Y) \le \frac{k\mu}{\eta r_{\mathrm{x}}} \text{ or } \lambda_{m+1}(P_{\tilde{X}}Y) > \frac{k\mu}{\eta r_{\mathrm{x}}} \Big\}.$$

Note that $P_{\tilde{X}}Y = \tilde{X}B + P_{\tilde{X}}\varepsilon$, which yields $\lambda_1(P_{\tilde{X}}\varepsilon) \ge \lambda_m(\tilde{X}B) - \lambda_m(P_{\tilde{X}}Y)$ and $\lambda_1(P_{\tilde{X}}\varepsilon) \ge \lambda_{m+1}(P_{\tilde{X}}Y) - \lambda_{m+1}(\tilde{X}B)$. Therefore, $\lambda_m(P_{\tilde{X}}Y) \le \dfrac{k\mu}{\eta r_{\mathrm{x}}}$ implies $\lambda_1(P_{\tilde{X}}\varepsilon) \ge \lambda_m(\tilde{X}B) - \dfrac{k\mu}{\eta r_{\mathrm{x}}}$, while $\lambda_{m+1}(P_{\tilde{X}}Y) > \dfrac{k\mu}{\eta r_{\mathrm{x}}}$ implies $\lambda_1(P_{\tilde{X}}\varepsilon) \ge \dfrac{k\mu}{\eta r_{\mathrm{x}}} - \lambda_{m+1}(\tilde{X}B)$. We thus have

$$P(\tilde{r} \ne m) \le P\Big(\lambda_1(P_{\tilde{X}}\varepsilon) \ge \min\Big\{ \lambda_m(\tilde{X}B) - \frac{k\mu}{\eta r_{\mathrm{x}}},\ \frac{k\mu}{\eta r_{\mathrm{x}}} - \lambda_{m+1}(\tilde{X}B) \Big\} \Big).$$

In addition, by the assumed conditions on $\lambda_{m+1}(\tilde{X}B)$ and $\lambda_m(\tilde{X}B)$, the proof of the lemma gets completed. □

Lemma 3. *Let $\{M_l\}$ be a finite sequence of matrices with dimension $n \times r$, and $\{\xi_l\}$ be a finite sequence of independent standard normal variables. Consider the matrix Gaussian series $Z = \sum_l \xi_l M_l$, and let $V(Z)$ be the matrix variance statistic of the sum, that is,*
$V(Z) = \max\{||E(ZZ^T)||_2, ||E(Z^TZ)||_2\} = \max\{||\sum_l M_l M_l^T||_2, ||\sum_l M_l^T M_l||_2\}$.
Then,

$$E||Z||_2 \leq \sqrt{2V(Z)\log(n+r)}.$$

Moreover, for all $t_0 \geq 0$, we have

$$P\{||Z||_2 \geq t_0\} \leq (n+r)\exp\left(\frac{-t_0^2}{2V(Z)}\right)$$

and

$$P\{||Z||_2 \geq E||Z||_2 + t_0\} \leq \exp\left(\frac{-t_0^2}{2V(Z)}\right).$$

Proof. The proof of Lemma 3 can be found in Chapter 4 of Tropp [23]. □

Lemma 4. *Let the SVD of X be as in (8). For a fixed $l \geq 0$, suppose S_1 has full row rank. Then, the approximation error satisfies*

$$\| (I - P_{XS})X \|_F^2 \leq \| \Lambda_2 \|_F^2 + \| \Lambda_2 S_2 S_1^+ \|_F^2 . \tag{29}$$

Lemma 5. *If the matrices M and N are fixed, and a standard Gaussian matrix G is drawn, then*

$$E[\| MGN \|_F^2] = \| M \|_F^2 \| N \|_F^2 . \tag{30}$$

Lemma 6. *Suppose g is a Lipschitz function on matrices satisfying*

$$|g(X) - g(Y)| \leq L \| X - Y \|_F, \quad \textit{for all } X \textit{ and } Y, \tag{31}$$

where L denotes Lipschitz constant. Draw a standard Gaussian matrix G. Then,

$$P\{g(G) \geq Eg(G) + Lt\} \leq e^{-t^2/2}, \quad \textit{for all } t \geq 1. \tag{32}$$

Lemma 7. *Let G be a $l \times k$ standard Gaussian matrix, and $k \geq l + 4$. Then, for all $t \geq 1$,*

$$P\left\{ \| G^+ \|_F \geq t\sqrt{\frac{3l}{k-l+1}} \right\} \leq t^{-(k-l)}. \tag{33}$$

Proof. Proofs of Lemmas 4–7 can be found in Halko et al. [10]. □

Lemma 8. *Suppose $\tilde{X}$ and Λ_2 are as defined in (4) and (8), respectively, and $k \geq l + 4$, with l being a non-negative integer. Then, for all $\gamma \geq 1$, $t \geq 1$,*

$$\| X - \tilde{X} \|_F^2 \leq \left[\frac{3lt^2}{k-l+1}(\gamma+1)^2 + 1 \right] \| \Lambda_2 \|_F^2$$

with failure probability at most $e^{-\gamma^2/2} + t^{-(k-l)}$.

Proof. Let $g(X) = \| \Lambda_2 X S_1^+ \|_F$. Then, by using triangle inequality of norm and some norm properties, we have

$$\begin{aligned} |g(X) - g(Y)| &= | \| \Lambda_2 X S_1^+ \|_F - \| \Lambda_2 Y S_1^+ \|_F | \\ &\leq \| \Lambda_2 (X - Y) S_1^+ \|_F \\ &\leq \| \Lambda_2 \|_F \| S_1^+ \|_F \| (X - Y) \|_F . \end{aligned}$$

This implies, by the Lipschitz constant, that

$$L \leq \| \Lambda_2 \|_F \| S_1^+ \|_F . \tag{34}$$

By the properties of expected values and Lemma 5, we then obtain

$$\left[E[\| \Lambda_2 S_2 S_1^+ \|_F | S_1] \right]^2 \leq E[\| \Lambda_2 S_2 S_1^+ \|_F | S_1]^2 = \| \Lambda_2 \|_F^2 \| S_1^+ \|_F^2;$$

that is,

$$E[\| \Lambda_2 S_2 S_1^+ \|_F | S_1] \leq \| \Lambda_2 \|_F \| S_1^+ \|_F . \tag{35}$$

We now define the event

$$E_p = \left\{ S_1 : \ \| S_1^+ \|_F \leq t \sqrt{\frac{3l}{k - l + 1}} \right\}, \quad for\ all\ t \geq 1.$$

Combining (32), (34) and (35), we have

$$P\left\{ \| \Lambda_2 S_2 S_1^+ \|_F | E_p \geq \| \Lambda_2 \|_F \| S_1^+ \|_F | E_p + \gamma \| \Lambda_2 \|_F \| S_1^+ \|_F | E_p \right\} \leq e^{-\gamma^2/2},$$

for all $\gamma \geq 1$, which is equivalent to

$$P\left\{ \| \Lambda_2 S_2 S_1^+ \|_F^2 | E_p \geq \| \Lambda_2 \|_F^2 \frac{3lt^2}{k - l + 1} (1 + \gamma)^2 \right\} \leq e^{-\gamma^2/2},$$

for all γ and $t \geq 1$.

Moreover, by the nature of probability and the use of Lemma 7 have

$$\begin{aligned} & P\left\{ \| \Lambda_2 S_2 S_1^+ \|_F^2 \geq \| \Lambda_2 \|_F^2 \frac{3lt^2}{k - l + 1} (1 + \gamma)^2 \right\} \\ & = P\left\{ \left(\| \Lambda_2 S_2 S_1^+ \|_F^2 \geq \| \Lambda_2 \|_F^2 \frac{3lt^2}{k - l + 1} (1 + \gamma)^2 \right) E_p \right\} \\ & + P\left\{ \left(\| \Lambda_2 S_2 S_1^+ \|_F^2 \geq \| \Lambda_2 \|_F^2 \frac{3lt^2}{k - l + 1} (1 + \gamma)^2 \right) E_p^c \right\} \\ & \leq P\left\{ \| \Lambda_2 S_2 S_1^+ \|_F^2 | E_p \geq \| \Lambda_2 \|_F^2 \frac{3lt^2}{k - l + 1} (1 + \gamma)^2 \right\} + P(E_p^c) \\ & \leq e^{-\gamma^2/2} + t^{-(k-l)}. \end{aligned}$$

Thus,

$$\begin{aligned} & P\left\{ \| \Lambda_2 S_2 S_1^+ \|_F^2 + \| \Lambda_2 \|_F^2 \geq \| \Lambda_2 \|_F^2 \left[\frac{3lt^2}{k - l + 1} (1 + \gamma)^2 + 1 \right] \right\} \\ & \leq e^{-\gamma^2/2} + t^{-(k-l)}. \end{aligned} \tag{36}$$

Upon using Lemma 4 and the fact $P_{XS} = QQ^T$, the proof gets completed. □

Appendix B: Proofs of main theorems and corollaries

B.1. Proof of Theorem 2.1

Proof. By the given condition that $\lambda_{r_0}(\tilde{X}B) > \dfrac{2k\mu}{\eta r_{\mathrm{x}}}$, we have $\lambda_{r_0}(\tilde{X}B) > \dfrac{(1+\delta)k\mu}{\eta r_{\mathrm{x}}}$. Note that $\lambda_{r_0+1}(\tilde{X}B) = 0$, and so $\lambda_{r_0+1}(\tilde{X}B) < \dfrac{(1-\delta)k\mu}{\eta r_{\mathrm{x}}}$. For this proof, we express ε as

$$\varepsilon = \sum_{i=1}^{n}\sum_{j=1}^{r}\xi_{ij}\Sigma_{ij},$$

where $\{\xi_{ij}\}$ is a sequence of independent standard normal variables, and Σ_{ij} is a $n \times r$ matrix with the (ij)th entry being σ_{ij} and remaining entries being 0. Let $M_{ij} = P_{\tilde{X}}\Sigma_{ij}$. Thus,

$$P_{\tilde{X}}\varepsilon = \sum_{i=1}^{n}\sum_{j=1}^{r}\xi_{ij}P_{\tilde{X}}\Sigma_{ij} = \sum_{i=1}^{n}\sum_{j=1}^{r}\xi_{ij}M_{ij}$$

Further, by Lemmas 2 and 3, and let $t_0 = \theta\sqrt{2V(P_{\tilde{X}}\varepsilon)\log(n+r)}$, we have

$$\begin{aligned}
P(\tilde{r} \neq r_0) &\leq P\Big\{\lambda_1(P_{\tilde{X}}\varepsilon) \geq \frac{\delta k\mu}{\eta r_{\mathrm{x}}}\Big\} \\
&= P\Big\{\lambda_1(P_{\tilde{X}}\varepsilon) \geq (1+\theta)\sqrt{2V(P_{\tilde{X}}\varepsilon)\log(n+r)}\Big\} \\
&\leq P\{\lambda_1(P_{\tilde{X}}\varepsilon) \geq E[\lambda_1(P_{\tilde{X}}\varepsilon)] + t_0\} \\
&\leq \exp\Big(\frac{-\theta^2 2V(P_{\tilde{X}}\varepsilon)\log(n+r)}{2V(P_{\tilde{X}}\varepsilon)}\Big) \\
&= \exp\big(-\theta^2\log(n+r)\big) \longrightarrow 0 \text{ as } n+r \longrightarrow \infty.
\end{aligned}$$

Thus, the proof of the theorem gets completed. □

B.2. Proof of Theorem 2.2

Proof. For any $p \times r$ matrix C, by the definition of $\tilde{B}$, we have

$$\| Y - \tilde{X}\tilde{B} \|_F^2 + 2\mu \| \tilde{X}\tilde{B} \|_* \leq \| Y - \tilde{X}C \|_F^2 + 2\mu \| \tilde{X}C \|_* .$$

Recall that

$$\begin{aligned}
\| Y - \tilde{X}\tilde{B} \|_F^2 &= \| XB + \varepsilon - \tilde{X}\tilde{B} \|_F^2 \\
&= \| XB - \tilde{X}\tilde{B} \|_F^2 + \| \varepsilon \|_F^2 + 2 < \varepsilon,\ XB - \tilde{X}\tilde{B} >_F
\end{aligned}$$

and

$$\| Y - \tilde{X}C \|_F^2 = \| XB + \varepsilon - \tilde{X}C \|_F^2$$

$$= \quad \| XB - \tilde{X}C \|_F^2 + \| \varepsilon \|_F^2 + 2 < \varepsilon,\ XB - \tilde{X}C >_F .$$

These imply that

$$\begin{aligned} \| XB - \tilde{X}\tilde{B} \|_F^2 \quad &\leq \quad \| XB - \tilde{X}C \|_F^2 + 2 < \varepsilon,\ \tilde{X}\tilde{B} - \tilde{X}C >_F \\ &\quad + 2\mu\big(\| \tilde{X}C \|_* - \| \tilde{X}\tilde{B} \|_* \big). \end{aligned} \tag{37}$$

From the facts that $P_{\tilde{X}}\tilde{X} = \tilde{X}$, $< M,\ N >_F \leq ||M||_2||N||_*$ and $||N||_* \leq \sqrt{r(N)}||N||_F$, we have

$$\begin{aligned} < \varepsilon,\ \tilde{X}\tilde{B} - \tilde{X}C >_F \quad &= \quad < P_{\tilde{X}}\varepsilon,\ \tilde{X}\tilde{B} - \tilde{X}C >_F \\ &\leq \quad \| P_{\tilde{X}}\varepsilon \|_2 \| \tilde{X}\tilde{B} - \tilde{X}C \|_* \\ &\leq \quad \sqrt{2r_0} \| P_{\tilde{X}}\varepsilon \|_2 \| \tilde{X}\tilde{B} - \tilde{X}C \|_F . \end{aligned} \tag{38}$$

Moreover, using the triangle inequality of norm, we have

$$\mu\big(\| \tilde{X}C \|_* - \| \tilde{X}\tilde{B} \|_* \big) \leq \mu \| \tilde{X}\tilde{B} - \tilde{X}C \|_* \leq \mu\sqrt{2r_0} \| \tilde{X}\tilde{B} - \tilde{X}C \|_F . \tag{39}$$

Now, combining (37), (38) and (39), we have

$$\begin{aligned} \| XB - \tilde{X}\tilde{B} \|_F^2 \quad &\leq \quad \| XB - \tilde{X}C \|_F^2 + 2\sqrt{2r_0} \| \tilde{X}\tilde{B} - \tilde{X}C \|_F \big(\| P_{\tilde{X}}\varepsilon \|_2 + \mu \big) \\ &\leq \quad \| XB - \tilde{X}C \|_F^2 + 2\sqrt{2r_0}\big(\| XB - \tilde{X}\tilde{B} \|_F \\ &+ \quad \| XB - \tilde{X}C \|_F \big)\big(\| P_{\tilde{X}}\varepsilon \|_2 + \mu \big), \end{aligned}$$

and it then follows that

$$\| XB - \tilde{X}\tilde{B} \|_F \leq \| XB - \tilde{X}C \|_F + 2\sqrt{2r_0}\big(\| P_{\tilde{X}}\varepsilon \|_2 + \mu \big). \tag{40}$$

As shown in the proof of Theorem 2.1,

$$P\{ \| P_{\tilde{X}}\varepsilon \|_2 \geq \delta\mu \} \leq \exp\big(-\theta^2 \log(n + r) \big).$$

Thus, we obtain

$$\| XB - \tilde{X}\tilde{B} \|_F \leq \| XB - \tilde{X}C \|_F + 2\mu(1 + \delta)\sqrt{2r_0} \tag{41}$$

with failure probability at most $\exp\big(-\theta^2 \log(n + r) \big)$.

Further, setting $C = B$, we have

$$\| XB - \tilde{X}C \|_F \leq \| X - \tilde{X} \|_F \| B \|_F . \tag{42}$$

Upon combining (41) and (42), and using Lemma 8, the proof gets completed. □

B.3. Proof of Theorem 2.3

Proof. From (40) and the fact that $\| A \|_2 \leq \| A \|_F$ for any matrix A, we obtain

$$E[\| XB - \tilde{X}\tilde{B} \|_F |Q] \leq \| XB - \tilde{X}C \|_F + 2\sqrt{2r_0}\Big\{E[\| P_{\tilde{X}}\varepsilon \|_F |Q] + \mu\Big\}. \quad (43)$$

Recall that if a function g is concave, then $E[g(X)] \leq g[E(X)]$, and so we have

$$E[\| P_{\tilde{X}}\varepsilon \|_F |Q] = E\sqrt{tr(\varepsilon^T P_{\tilde{X}}\varepsilon)|Q} \leq \sqrt{tr[E(\varepsilon\varepsilon^T)P_{\tilde{X}}]}. \quad (44)$$

Further, by the law of iterated expectations and the fact that $P_{\tilde{X}} = QQ^T$, we have

$$E[\| P_{\tilde{X}}\varepsilon \|_F] = E\Big\{E[\| P_{\tilde{X}}\varepsilon \|_F |Q]\Big\} \leq \sqrt{\frac{k}{n}tr[E(\varepsilon\varepsilon^T)]} \quad (45)$$

and

$$E_Q[\| XB - \tilde{X}C \|_F] \leq \Big[\| XB \|_F^2 - \frac{2k}{n} < XC,\ XB >_F + \frac{k}{n} \| XC \|_F^2 \Big]^{1/2}. \quad (46)$$

Note that $\varepsilon = [\varepsilon_{ij}]_{n\times r}$ is an error matrix and the entries are independent of each other with mean zero and variance σ_{ij}^2, and so

$$\sqrt{\frac{k}{n}tr[E(\varepsilon\varepsilon^T)]} = \Big[\frac{k}{n}\sum_{i=1}^{n}\sum_{j=1}^{r}\sigma_{ij}^2\Big]^{1/2}. \quad (47)$$

Further, setting $C = B$, we obtain

$$\begin{aligned} E_Q[\| XB - \tilde{X}B \|_F] &\leq \Big(1 - \frac{k}{n}\Big)^{1/2} \| XB \|_F \\ &\leq \Big(1 - \frac{k}{n}\Big)^{1/2}\Big[\sum_{i=1}^{q}\lambda_i^2(X)\Big]^{1/2} \| B \|_F . \end{aligned} \quad (48)$$

Upon combining (43)–(48), we complete the proof of the theorem. □

B.4. Proof of Corollary 2.4

Proof. From (40) and the proof of Theorem 2.3, we have

$$E[\| XB - X\hat{B}_{NC} \|_F] \leq 2\sqrt{2r_0}\big(\sqrt{E[tr(\varepsilon\varepsilon^T P_X)]} + \mu\big). \quad (49)$$

Note that $X = \Gamma\Lambda P^T$ and $X^+ = P\Lambda^{-1}\Gamma^T$, and so $P_X = X(X^TX)^+X^T = XX^+ = \Gamma\Gamma^T$. Then,

$$\sqrt{E[tr(\varepsilon\varepsilon^T P_X)]} = \Big[\sum_{i=1}^{n}\Big(\sum_{j=1}^{r}\sigma_{ij}^2\Big)\Big(\sum_{h=1}^{q}\gamma_{ih}^2\Big)\Big]^{1/2}. \quad (50)$$

By combining (49) and (50), we complete the proof. □

B.5. Proof of Corollary 2.5

Proof. Using (40), with failure probability at most $\exp\big(-\theta^2\log(n+r)\big)$, we obtain

$$\begin{aligned} \| XB - \hat{X}\hat{B} \|_F &\leq \| XB - \hat{X}B \|_F + 2\mu(1+\delta)\sqrt{2r_0} \\ &\leq \| X - \hat{X} \|_F \| B \|_F + 2\mu(1+\delta)\sqrt{2r_0}. \end{aligned} \tag{51}$$

Moreover, by (8) and (14), we obtain

$$\begin{aligned} X - \hat{X} &= \Gamma\Lambda P^T(I_p - P_1P_1^T) \\ &= (\Gamma_1, \quad \Gamma_2)\begin{pmatrix}\Lambda_1 & \\ & \Lambda_2\end{pmatrix}\begin{pmatrix}P_1^T \\ P_2^T\end{pmatrix}P_2P_2^T \\ &= \Gamma_2\Lambda_2P_2^T. \end{aligned}$$

Hence, we have

$$\| X - \hat{X} \|_F = \sqrt{tr(P_2\Lambda_2\Gamma_2^T\Gamma_2\Lambda_2P_2^T)} = \sqrt{tr(\Lambda_2^2)} = \Big[\sum_{i=k+1}^{q}\lambda_i^2(X)\Big]^{1/2}. \tag{52}$$

Upon combining (51) and (52), we complete the proof. □

B.6. Proof of Corollary 2.6

Proof. From (40) and the proof of Theorem 2.3, we have

$$E[\| XB - \hat{X}\hat{B} \|_F] \leq E[\| XB - \hat{X}B \|_F] + 2\sqrt{2r_0}\big(\sqrt{E[tr(\varepsilon\varepsilon^TP_{\hat{X}})]} + \mu\big). \tag{53}$$

Note that $X = \Gamma\Lambda P^T = \Gamma_1\Lambda_1P_1^T + \Gamma_2\Lambda_2P_2^T$, and $X^+ = P\Lambda^{-1}\Gamma^T = P_1\Lambda_1^{-1}\Gamma_1^T + P_2\Lambda_2^{-1}\Gamma_2^T$, and so $P_{\hat{X}} = \hat{X}(\hat{X}^T\hat{X})^+\hat{X}^T = XP_1P_1^TX^+ = \Gamma_1\Gamma_1^T$. Hence, we obtain

$$\sqrt{E[tr(\varepsilon\varepsilon^TP_{\hat{X}})]} = \Big[\sum_{i=1}^{n}\Big(\sum_{j=1}^{r}\sigma_{ij}^2\Big)\Big(\sum_{h=1}^{k}\gamma_{ih}^2\Big)\Big]^{1/2}. \tag{54}$$

Now, upon combining (53) and (54), the proof gets completed. □

Acknowledgments

We express our sincere thanks to the Editor, the Associate Editor and the reviewer for their incisive comments and suggestions on an earlier version of this manuscript which led to this much improved version.

References

[1] Ahfock, D., Astle, J.W. and Richardson, S. (2021). Statistical properties of sketching algorithms. *Biometrika* 108: 283–297. MR4259132

[2] Anderson, T.W. (1999). Asymptotic distribution of the reduced rank regression estimator under general conditions. *The Annals of Statistics*, 27: 1141–1154. MR1740118

[3] Breiman, L.(1996). Bagging predictors. *Machine Learning*, 24: 123–140.

[4] Bunea, F., She, Y. and Wegkamp, M. (2011). Optimal selection of reduced rank estimators of high-dimensional matrices. *The Annals of Statistics*, 39: 1282–1309. MR2816355

[5] Cannings, T.I. and Samworth, R.J. (2017). Random-projection ensemble classification. *Journal of the Royal Statistical Society: Series B*, 79: 959–1035. MR3689307

[6] Chen, K., Dong, H. and Chan, K.S. (2013). Reduced rank regression via adaptive nuclear norm penalization. *Biometrika*, 100: 901–920. MR3142340

[7] Chin, K., DeVries, S., Fridlyand, J., Spellman, P., Roydasgupta, R., Kuo,W.-L., Lapuk, A., Neve, R., Qian, Z., Ryder, T., Chen, F., Feiler, H., Tokuyasu, T., Kingsley, C., Dairkee, S., Meng, Z., Chew, K., Pinkel, D., Jain, A., Ljung, B., Esserman, L., Albertson, D.,Waldman, F. and Gray, J. (2006). Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. *Cancer Cell*, 10: 529–541.

[8] Dobriban, E. and Liu, S. (2019). Asymptotics for sketching in least squares regression. arXiv:1810.06089v2.

[9] Fern, X.Z. and Brodley, C.E. (2003). Random projection for high dimensional data clustering: A cluster ensemble approach. In: *Proceedings of the 20th International Conference on Machine Learning (ICML-2003)*, pages 186–193. MR

[10] Halko, N., Martinsson, P.G. and Tropp, J.A. (2011). Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53: 217–288. MR2806637

[11] Horn, S.D., Horn, R.A. and Duncan, D.B. (1975). Estimating heteroscedastic variances in linear models. *Journal of the American Statistical Association*, 70: 380–385. MR0370946

[12] Mirsky, L. (1975). A trace inequality of John von Neumann. *Monatshefte für Mathematik*, 79: 303–306. MR0371930

[13] Mukherjee, A. and Zhu, J. (2011). Reduced rank ridge regression and its kernel extensions. *Statistical Analysis and Data Mining*, 4: 612–622. MR2758084

[14] Peng, J., Zhu, J., Bergamaschi, A., Han, W., Noh, D.Y., Pollack, J.R. and Wang, P. (2010). Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer. *Annals of Applied Statistics*, 4: 53–77. MR2758084

[15] Pilanci, M. and Wainwright, M.J. (2015). Randomized sketches of

convex programs with sharp guarantees. *IEEE Transactions on Information Theory*, 61: 5096–5115. MR3386504

[16] Pilanci, M. and Wainwright, M.J. (2016). Iterative Hessian sketch: Fast and accurate solution approximation for constrained least-squares. *The Journal of Machine Learning Research*, 17: 1842–1879. MR3504613

[17] Pilanci, M. and Wainwright, M.J. (2017). Newton sketch: A near linear-time optimization algorithm with linearquadratic convergence. *SIAM Journal on Optimization*, 27: 205–245. MR3612185

[18] Rao, C.R. (1970). Estimation of heteroscedastic variances in linear models. *Journal of the American Statistical Association*, 65: 161–172. MR0286221

[19] Reinsel, G.C. and Velu, R.P. (1998). *Multivariate Reduced-Rank Regression: Theory and Applications*. Springer, New York.

[20] Slawski, M. (2018). On principal components regression, random projections, and column subsampling. *Electronic Journal of Statistics*, 12: 3673–3712. MR3870509

[21] Stoica, P. and Viberg, M. (1996). Maximum likelihood parameter and rank estimation in reduced-rank multivariate linear regressions. *IEEE Transactions on Signal Processing*, 44: 3069–3079.

[22] Thanei, G.-A., Heinze, C. and Meinshausen, N. (2017). Random projections for large-scale regression. *InBig and complex data analysis 2017* pages 51–68. Springer, Cham. MR3644120

[23] Tropp, J.A. (2015). *An Introduction to Matrix Concentration Inequalities*. Now Pub.

[24] von Neumann, J. (1937). Some matrix inequalities and metrization of matrix-space. *Tomsk University Reviews*, 1: 286–300.

[25] Wang, S., Gittens, A. and Mahoney, M.W. (2018). Sketched ridge regression: Optimization perspective, statistical perspective, and model averaging. *Journal of Machine Learning Research*, 18: 1–50. MR3827106

[26] Witten, D.M., Tibshirani, R.J. and Hastie, T.J. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10: 515-534.

[27] 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. MR2323756