Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 October 2022 doi:10.20944/preprints202210.0169.v1

Article

Minimax lower bounds for high-dimensional multi-response errors-in-variables regression

Xin Li [1] and Dongya Wu [2,*]

[1] School of Mathematics, Northwest University, Xi'an, 710069, P. R. China; lixin@nwu.edu.cn
[2] School of Information Science and Technology, Northwest University, Xi'an, 710069, P. R. China; wudongya@nwu.edu.cn
* Correspondence: wudongya@nwu.edu.cn

Abstract: Noisy data is always encountered in real applications, such as bioinformatics, neuroimage and remote sensing. Existing methods mainly consider linear or generalized linear errors-in-variables regression, while relatively little attention is paid for the multivariate response case, and how to evaluate the estimation performance under perturbed covariates is still an open question. In this paper, we consider the information-theoretic limitations of estimating a low-rank matrix in the multi-response errors-in-variables regression model. By application of the information theory and statistical techniques on concentration inequalities, the minimax lower bound is provided in terms of the squared Frobenius loss, which recaptures the rate provided under the clean covariate assumption in previous literatures. Hence our result further indicates that though under the more realistic errors-in-variables situation, no more samples are required so as to achieve a rate-optimal estimation.

Keywords: low-rank matrices; errors-in-variables models; lower bounds; Kullback-Leibler divergence; information-theoretic limitations

1. Introduction

Recent decades have witnessed a fruitful results on high-dimensional statistics, including theory and application; see the books [1,2] for an overall review. In order to deal with the curse of dimensionality, different statistical models have been proposed with certain low-dimensional structures, such as sparse linear regression, low-rank matrix regression and so on. Specifically, the multi-response regression model, which is an important instance of matrix regression, has been deeply investigated in theoretical aspects [3,4] and widely used in real applications such as neuroimage analysis [5,6]. Consider the following multi-response regression model

$$Y = X\Theta^* + \epsilon, \tag{1}$$

where $\Theta^* \in \mathbb{R}^{d_1 \times d_2}$ is the unknown underlying parameter matrix, $Y \in \mathbb{R}^{n \times d_2}$ is the response matrix, $X \in \mathbb{R}^{n \times d_1}$ is the covariate matrix, and $\epsilon \in \mathbb{R}^{n \times d_2}$ is the noise matrix. The parameter matrix Θ^* is usually endowed with some structured constraints such as low-rankness in order to achieve consistent estimation.

There are mainly two research directions in the field of high-dimensional statistics. On the one hand, researchers seek to construct estimators with fast convergence rates. On the other hand, it is also important to understand the fundamental or information-theoretic limitations of any estimator in order to evaluate the corresponding performance. The former goal can be achieved by establishing upper bounds on estimation errors via statistical techniques such as concentration inequalities, while the latter one requires tools from information theory to derive lower bounds on some quantitative criteria.

Given an arbitrary estimator of the true unknown parameter, many criteria can be utilized to evaluate the quality of the estimate. From the decision-theoretic framework, it is typical to introduce a loss function which represents the loss induced by the estimating procedure. Then in the minimax formalism, a worst-case loss function is constructed and then minimized to characterize the optimal rate. Inequalities for hypothesis test and Fano's inequality are often used to derive lower bounds for the worst-case loss with the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 October 2022 doi:10.20944/preprints202210.0169.v1

help of estimating some information-theoretic quantities such as the mutual information, Kullback–Leibler (KL) divergence, and total variation distance; see, e.g., [7,8].

Note that when $d_2 = 1$ and the underlying parameter Θ^* is endowed with vector sparsity, model (1) reduces to sparse linear regression and attracts a lot of works focusing on the minimax estimation. For example, Ye and Zhang [9] and Raskutti *et al.* [10] provided lower bounds on minimax rates of convergence for estimation via standard information-theoretic techniques, respectively; Wang et. al. [11] established the minimax optimal rates of convergence and constructed an adaptive optimal estimator by virtue of the proposed aggregation strategy. For the low-rank problem, researchers mainly focused on the matrix completion problem and established the minimax optimal rate; see, e.g., [12–14]. Ja and Kutzarova [15] considered general low-rank matrix recovery problems and provided the worst-case error utilizing the Gelfand widths of certain identity mappings between finite-dimensional Schatten spaces.

All of the above mentioned works were based on the assumption that the covariates are collected with clean data, which is standard in theoretical analysis. However, this assumption may always be violated in real world problems due to instrumental constraints or a lack of observations. This is to say that the obtained covariates are usually perturbed with certain measurement error. What if we ignore the measurement error and naively apply methods for clean covariates in this noisy case? The answer is depressing since it has been shown in [16] by simulations that this operation can only lead to misleading results. Therefore, it is more necessary and realistic to investigate statistical models in which only perturbed observations of true covariates are observed.

Recently, researchers began to devote to errors-in-variables regression problems and most of the results were established for statistical inference of linear or generalized linear regression; see, e.g., [17–20] and references therein. In the information-theoretic aspect, Loh and Wainwright [21] and Li and Wu [22] considered linear errors-in-variables regression and established the minimax lower bound for estimating a sparse vector via calculating the corresponding KL divergence over certain sparse sets for vectors, respectively.

However, until now, relatively little attention is paid for multivariate regression model (1) with measurement error. Although it is a natural and simple idea to vectorize both the coefficient and response matrices to reduce the original multivariate model to the univariate response one, so that the above mentioned results can be applied directly. Unfortunately, the low-rankness of a matrix is rather different from the sparsity of a vector due to the more sophisticated manifold structure [23]. Moreover, the multivariate nature of the responses enables one to build more complex models for modern large-scale association analysis, such as fMRI image analyses [6] and physiological network analyses [24], and thus has a substantial wider application than that of the univariate model.

In this paper, we study the information-theoretic limitations for multi-response errors-in-variables regression. The estimation on the lower bound is first reduced to a multiway hypothesis testing problem, and then Fano's inequality [8] is applied to lower bound the error probability. The main contributions of this paper are as follows. First, the KL divergence involved in mutual information is estimated with the help of a concentration inequality on random matrix multiplication. Then we establish lower bounds on the minimax loss function in terms of the squared Frobenius norm for a certain class of low-rank matrices. This lower bound agrees with the upper bound given before in our another work [25] up to constant factors, implying the optimality of the proposed estimator therein. Moreover, the minimax lower bound recaptures the rate provided under the clean covariate assumption in previous literatures [12,14,15], a result that further indicates that though in the more realistic errors-in-variables situation, no more samples are required so as to achieve a rate-optimal estimation.

The remainder of this paper is organized as follows. In Section 2, we provide background on the multi-response errors-in-variables regression model and minimax estimation problems. In Section 3, we establish our main results on lower bounds on minimax estimation. Conclusions and future work are discussed in Section 4.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 October 2022 doi:10.20944/preprints202210.0169.v1

We end this section by introducing some notations for future reference. All vectors are column vectors following classical mathematical convention. For $d \geq 1$, let $\mathbb{I}_d$ stand for the $d \times d$ identity matrix. For a matrix $X \in \mathbb{R}^{n\times d}$, let X_{ij} $(i = 1, \ldots, n, j = 1, 2, \cdots, d)$ denote its ij-th entry, $X_{i\cdot}$ $(i = 1, \ldots, n)$ denote its i-th row, $X_{\cdot j}$ $(j = 1, 2, \cdots, d)$ denote its j-th column. When X is a square matrix, i.e., $n = d$, we use $\mathrm{diag}(X)$ stand for the diagonal matrix with its diagonal elements equal to $X_{11}, X_{22}, \cdots, X_{dd}$. We write $\lambda_{\min}(X)$ and $\lambda_{\max}(X)$ to denote the minimal and maximum eigenvalues of a matrix X, respectively. For a matrix $\Theta \in \mathbb{R}^{d_1\times d_2}$, define $d = \min\{d_1, d_2\}$, and denote its singular values in decreasing order by $\sigma_1(\Theta) \geq \sigma_2(\Theta) \geq \cdots \sigma_d(\Theta) \geq 0$. We use $|||\cdot|||$ to denote different types of matrix norms based on singular values, including the nuclear norm $|||\Theta|||_* = \sum_{j=1}^{d} \sigma_j(\Theta)$, the spectral or operator norm $|||\Theta|||_{\mathrm{op}} = \sigma_1(\Theta)$, and the Frobenius norm $|||\Theta|||_F = \sqrt{\mathrm{trace}(\Theta^\top\Theta)} = \sqrt{\sum_{j=1}^{d} \sigma_j^2(\Theta)}$.

2. Problem setup

In this section, we provide a detailed description of the multi-response errors-in-variables regression model and then the minimax estimation problem.

Consider the high-dimensional multi-response regression model which represents the relationship between the response vector $Y_{i\cdot} \in \mathbb{R}^{d_2}$ and the covariate vector $X_{i\cdot} \in \mathbb{R}^{d_1}$

$$Y_{i\cdot} = \Theta^{*\top} X_{i\cdot} + \epsilon_{i\cdot} \qquad \text{for } i = 1, 2, \cdots, n, \tag{2}$$

where $\Theta^* \in \mathbb{R}^{d_1\times d_2}$ is the unknown parameter matrix, and $\epsilon_{i\cdot} \in \mathbb{R}^{d_2}$ is the observation noise of the response vector independent of $X_{j\cdot}$ $(\forall\ i, j)$. Model (1) can be expressed in a more compact matrix form. More precisely, define the multi-response matrix $Y = (Y_{1\cdot}, Y_{2\cdot}, \cdots, Y_{n\cdot})^\top \in \mathbb{R}^{n\times d_2}$ with similar definitions for the covariate matrix $X \in \mathbb{R}^{n\times d_1}$ and the noise matrix $\epsilon \in \mathbb{R}^{n\times d_2}$ by the corresponding vectors $\{X_{i\cdot}\}_{i=1}^n$ and $\{\epsilon_{i\cdot}\}_{i=1}^n$, respectively. Then model (2) is re-written as

$$Y = X\Theta^* + \epsilon. \tag{3}$$

We work within the high-dimensional scenario in which the number of covariates or responses (i.e., d_1 or d_2) may be possibly more than the sample size n. It is well known that consistent estimation cannot be achieved under this high-dimensional regime unless the parameter space is imposed with additional low-dimensional structures, such as low-rankness in matrix estimation problems. Particularly, assume $R_0 \ll \min\{d_1, d_2\}$, and we shall consider the following low-rank matrix set throughout this paper

$$\mathbb{V}_{d_1,d_2}^{R_0} := \left\{ \Theta \in \mathbb{R}^{d_1\times d_2} \,\middle|\, \Theta^\top\Theta = \begin{pmatrix} \mathbb{I}_{R_0} & 0 \\ 0 & 0 \end{pmatrix} \right\}. \tag{4}$$

Note that any matrix $\Theta \in \mathbb{V}_{d_1,d_2}^{R_0}$ is of $\mathrm{rank}(\Theta) = R_0$, and thus Θ is exact low-rank due to the assumption $R_0 \ll \min\{d_1, d_2\}$.

Recall the standard multi-response model (3), the covariate matrix X is assumed to be correctly observed. Unfortunately, in real applications, the covariates are usually perturbed with noise. In this more realistic situation, one can only observe the noisy matrix Z instead of the true covariate matrix X. Throughout this paper, we consider the following errors-in-variables model with additive noise: For each $i = 1, 2, \cdots, n$, we observe $Z_{i\cdot} = X_{i\cdot} + W_{i\cdot}$, where $W_{i\cdot} \in \mathbb{R}^{d_1}$ is a random vector independent of $X_{i\cdot}$ with mean 0 and known covariance matrix Σ_w. When the noise covariance Σ_w is unknown, one can try to estimate it from the observed data by virtue of statistical techniques; see, e.g., [26]. For instance, a simple method is to estimate Σ_w from blank control observations which are independent of the noise. Concretely speaking, suppose that a matrix $W_0 \in \mathbb{R}^{n\times d_1}$ is observed independently with n i.i.d. vectors of measurement errors. Then the matrix $\frac{1}{n}W_0^\top W_0$ can be used as the estimate of Σ_w. Other sophisticated variant based on this method are also discussed in [26].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 October 2022 doi:10.20944/preprints202210.0169.v1

Throughout this paper, we impose a Gaussian random assumption on the model, that is, for $i = 1, 2, \cdots, n$, the vectors $X_{i\cdot}$, $W_{i\cdot}$ and $\epsilon_{i\cdot}$ are Gaussian with mean 0 and covariance matrices $\sigma_x^2 \mathbb{I}_{d_1}$, $\sigma_w^2 \mathbb{I}_{d_1}$ and $\sigma_\epsilon^2 \mathbb{I}_{d_2}$, respectively. For the sake of simplicity, we shall write $\sigma_z^2 = \sigma_x^2 + \sigma_w^2$.

Statistically, in order to estimate the underlying parameter Θ^*, researchers seek to construct an estimator $\hat{\Theta} : \mathbb{R}^{n\times d_1} \times \mathbb{R}^{n\times d_2} \to \mathbb{R}^{d_1\times d_2}$, which is a measurable function of the observed data (Z, Y). Then the information-theoretic task is to evaluate the estimation performance of $\hat{\Theta}$. It is standard to introduce a loss function $\mathcal{L}(\hat{\Theta}, \Theta^*)$, which expresses the loss induced by the estimator $\hat{\Theta}$ when the true parameter belongs to some certain set, that is, $\Theta^* \in \mathbb{V}_{d_1,d_2}^{R_0}$ in this paper. In the minimax formalism, we aim to lower bound the following worst-case loss in terms of the squared Frobenius norm

$$\mathcal{M}(\mathbb{V}_{d_1,d_2}^{R_0}) := \inf_{\hat{\Theta}} \sup_{\Theta^* \in \mathbb{V}_{d_1,d_2}^{R_0}} |||\hat{\Theta} - \Theta^*|||_F^2, \tag{5}$$

where the infimum is taken over all measurable functions $\hat{\Theta}$ of the collected data (Z, Y). Note that $\mathcal{M}(\mathbb{V}_{d_1,d_2}^{R_0})$ is stochastic because of the dependence of $\hat{\Theta}$ on the noise W and ϵ. Hence, lower bounds should be given in expectation or with high probability.

3. Main results

In this section, we provide the lower bound on the minimax risk which holds with high probability. This lower bound implies that the achievable upper bound established in [25, Theorem 1] is sharp and thus provide the information-theoretic foundation that no estimator can perform better than the nonconvex estimator proposed in [25] in the sense of statistical convergence rates.

Before the statement of Theorem 1, the following two lemmas are needed. The first one is statistical which reveals a concentration inequality on random matrix multiplication, while the second one is information-theoretical that tells us the the KL divergence between the distributions on the response variable Y induced by two different parameters $\Theta, \Theta' \in \mathbb{V}_{d_1,d_2}^{R_0}$. Recall that for two distributions $\mathbb{P}$ and $\mathbb{Q}$ with densities $d\mathbb{P}$ and $d\mathbb{Q}$ in regard to some base measure μ, the KL divergence is defined by $D(\mathbb{P}||\mathbb{Q}) = \int \log \frac{d\mathbb{P}}{d\mathbb{Q}} \mathbb{P}(d\mu)$. Let $\mathbb{P}_\Theta$ denote the distribution of Y in the multi-response regression model with additive errors, when Θ is given and Z is observed.

Lemma 1. *Let $t > 0$ be any constant, and $X \in \mathbb{R}^{n\times d_1}$ be a zero-mean sub-Gaussian matrix with parameters (Σ_x, σ_x^2). Then for any fixed matrix $\Theta \in \mathbb{R}^{d_1\times d_2}$, there exists a universal positive constant c_0 such that*

$$\mathbb{P}\left[\left|\frac{|||X\Theta|||_F^2}{n} - \mathbb{E}\left(\frac{|||X\Theta|||_F^2}{n}\right)\right| \geq t\right] \leq 2\exp\left(-c_0 n \min\left(\frac{t^2}{d_2^2\sigma_x^4}, \frac{t}{d_2\sigma_x^2}\right) + \log d_2\right).$$

Proof. By the definition of matrix Frobenius norm, one has that

$$\frac{|||X\Theta|||_F^2}{n} - \mathbb{E}\left(\frac{|||X\Theta|||_F^2}{n}\right) = \sum_{j=1}^{d_2}\left[\frac{\|X\Theta_{\cdot j}\|_2^2}{n} - \mathbb{E}\left(\frac{\|X\Theta_{\cdot j}\|_2^2}{n}\right)\right].$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 October 2022 doi:10.20944/preprints202210.0169.v1

Then it follows from elementary probability theory that

$$\begin{aligned}\mathbb{P}\left[\left|\frac{|||X\Theta|||_F^2}{n}-\mathbb{E}\left(\frac{|||X\Theta|||_F^2}{n}\right)\right|\le t\right] &= \mathbb{P}\left\{\left|\sum_{j=1}^{d_2}\left[\frac{\|X\Theta_{\cdot j}\|_2^2}{n}-\mathbb{E}\left(\frac{\|X\Theta_{\cdot j}\|_2^2}{n}\right)\right]\right|\le t\right\}\\ &\ge \mathbb{P}\left\{\bigcap_{j=1}^{d_2}\left\{\left|\frac{\|X\Theta_{\cdot j}\|_2^2}{n}-\mathbb{E}\left(\frac{\|X\Theta_{\cdot j}\|_2^2}{n}\right)\right|\le \frac{t}{d_2}\right\}\right\}\\ &\ge \sum_{j=1}^{d_2}\mathbb{P}\left[\left|\frac{\|X\Theta_{\cdot j}\|_2^2}{n}-\mathbb{E}\left(\frac{\|X\Theta_{\cdot j}\|_2^2}{n}\right)\right|\le \frac{t}{d_2}\right]-(d_2-1)\end{aligned}$$

On the other hand, note the assumption that X is a sub-Gaussian matrix with parameters (Σ_x,σ_x^2). Then [27, Lemma 14] is applicable to concluding that there exists a universal positive constant c_0 such that

$$\begin{aligned}\mathbb{P}\left[\left|\frac{|||X\Theta|||_F^2}{n}-\mathbb{E}\left(\frac{|||X\Theta|||_F^2}{n}\right)\right|\le t\right] &\ge d_2\left(1-2\exp\left(-c_0 n\min\left(\frac{t^2}{d_2^2\sigma_x^4},\frac{t}{d_2\sigma_x^2}\right)\right)\right)-(d_2-1)\\ &= 1-2\exp\left(-c_0 n\min\left(\frac{t^2}{d_2^2\sigma_x^4},\frac{t}{d_2\sigma_x^2}\right)+\log d_2\right),\end{aligned}$$

which completes the proof. □

Lemma 2. *In the additive noise setting, there exist universal positive constants (c_0,c_1) such that with probability at least $1-2\exp(-c_0 n+\log d_2)$, the KL divergence between the distributions of Y induced by any $\Theta,\Theta'\in\mathbb{V}_{d_1,d_2}^{R_0}$ is upper bounded as*

$$D(\mathbb{P}_\Theta||\mathbb{P}_{\Theta'})\le\frac{c_1 n\sigma_x^4}{\sigma_z^2\sigma_\epsilon^2}(|||\Theta-\Theta'|||_F^2+d_2). \tag{6}$$

Proof. For each $i=1,2,\cdots,n$ fixed, by the model setting, $(Y_{i\cdot},Z_{i\cdot})$ is jointly Gaussian with mean 0. Then by some elementary computations on the covariances, one has that

$$\begin{bmatrix}Y_{i\cdot}\\Z_{i\cdot}\end{bmatrix}\sim\mathcal{N}\left(\begin{bmatrix}0\\0\end{bmatrix},\begin{bmatrix}\Theta^\top\Sigma_x\Theta+\sigma_\epsilon^2\mathbb{I}_{d_2} & \Theta^\top\Sigma_x\\ \Sigma_x\Theta & \Sigma_x+\Sigma_w\end{bmatrix}\right).$$

Then it follows from standard results on the conditional distribution of Gaussian vectors that

$$Y_{i\cdot}|Z_{i\cdot}\sim\mathcal{N}(\Theta^\top\Sigma_x\Sigma_z^{-1}Z_{i\cdot},\Theta^\top(\Sigma_x-\Sigma_x\Sigma_z^{-1}\Sigma_x)\Theta+\sigma_\epsilon^2\mathbb{I}_{d_2}). \tag{7}$$

Now we can assume that σ_ϵ and σ_w are not both 0, since otherwise the conclusion follows trivially. For different parameters $\Theta,\Theta'\in\mathbb{V}_{d_1,d_2}^{R_0}$, define $\Sigma_\Theta:=\Theta^\top(\Sigma_x-\Sigma_x\Sigma_z^{-1}\Sigma_x)\Theta+\sigma_\epsilon^2\mathbb{I}_{d_2}$, and $\Sigma_{\Theta'}$ is given analogously. Recalling the previous assumptions that $\Sigma_w=\sigma_w^2\mathbb{I}_{d_1}$ and $\Sigma_\epsilon=\sigma_\epsilon^2\mathbb{I}_{d_2}$ and noting that $\Sigma_x=\sigma_x^2\mathbb{I}_{d_1}$, one has that $\Sigma_z=(\sigma_x^2+\sigma_w^2)\mathbb{I}_{d_1}$, and thus

$$\Sigma_\Theta=\left(\sigma_x^2-\frac{\sigma_x^4}{\sigma_z^2}\right)\Theta^\top\Theta+\sigma_\epsilon^2\mathbb{I}_{d_2}=\begin{pmatrix}\left(\frac{\sigma_x^2\sigma_w^2}{\sigma_z^2}+\sigma_\epsilon^2\right)\mathbb{I}_{R_0} & 0\\ 0 & \sigma_\epsilon^2\mathbb{I}_{d_2-R_0}\end{pmatrix}, \tag{8}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 October 2022 doi:10.20944/preprints202210.0169.v1

where the second equality is due to the fact that $\Theta \in \mathbb{V}_{d_1,d_2}^{R_0}$. By the independence of sampling, one has that $\mathbb{P}_\Theta$ is a product distribution of $Y_{i\cdot}|Z_{i\cdot}$ over all $i = 1, 2, \cdots, n$. Then it follows immediately from (7) that

$$\begin{aligned} D(\mathbb{P}_\Theta||\mathbb{P}_{\Theta'}) &= \mathbb{E}_{\mathbb{P}_\Theta}\left[\log\frac{\mathbb{P}_\Theta(Y)}{\mathbb{P}_{\Theta'}(Y)}\right] \\ &= \sum_{i=1}^{n}\Bigg[\frac{1}{2}\log\left(\frac{\det(\Sigma_{\Theta'})}{\det(\Sigma_\Theta)}\right) + \frac{1}{2}\left(\mathrm{tr}(\Sigma_{\Theta'}^{-1}\Sigma_\Theta) - d_2\right) \\ &\qquad + \frac{1}{2}((\Theta'-\Theta)^\top\Sigma_x\Sigma_z^{-1}Z_{i\cdot})^\top\Sigma_{\Theta'}^{-1}(\Theta'-\Theta)^\top\Sigma_x\Sigma_z^{-1}Z_{i\cdot})\Bigg] \\ &= \frac{n}{2}\log\left(\frac{\det(\Sigma_{\Theta'})}{\det(\Sigma_\Theta)}\right) + \frac{n}{2}\left(\mathrm{tr}(\Sigma_{\Theta'}^{-1}\Sigma_\Theta) - d_2\right) \\ &\qquad + \frac{1}{2}\mathrm{tr}((Z\Sigma_z^{-1}\Sigma_x(\Theta-\Theta'))^\top(Z\Sigma_z^{-1}\Sigma_x(\Theta-\Theta'))\Sigma_{\Theta'}^{-1}), \end{aligned} \tag{9}$$

By (8), one has that $\Sigma_\Theta = \Sigma'_\Theta$, and thus the first two terms in (9) are both equal to 0. It also follows from (8) that $\Sigma_{\Theta'}^{-1}$ is also diagonal with the first R_0 elements equal to $\frac{1}{\frac{\sigma_x^2\sigma_w^2}{\sigma_z^2}+\sigma_\epsilon^2}$ and the last $d_2 - R_0$ elements equal to $\frac{1}{\sigma_\epsilon^2}$. Since $\frac{1}{\frac{\sigma_x^2\sigma_w^2}{\sigma_z^2}+\sigma_\epsilon^2} \le \frac{1}{\sigma_\epsilon^2}$, combining these arguments with (9), we arrive at that

$$D(\mathbb{P}_\Theta||\mathbb{P}_{\Theta'}) \le \frac{1}{2\sigma_\epsilon^2}\left|\left|\left|Z\Sigma_z^{-1}\Sigma_x(\Theta-\Theta')\right|\right|\right|_F^2 = \frac{\sigma_x^4}{2\sigma_z^4\sigma_\epsilon^2}|||Z(\Theta-\Theta')|||_F^2.$$

This inequality, together with Lemma 1 ($t = d_2\sigma_z^2$), yields that there exist universal positive constants (c_0, c_1) such that (9) holds with probability at least $1 - 2\exp(-c_0 n + \log d_2)$. The proof is completed. □

Theorem 1. *In the additive noise setting, let $2 \le R_0 \le d_1 - R_0$. Then there exist universal positive constants (c_0, c_1) such that, with probability at least $1/2(1 - 2\exp(-c_0 n + \log d_2))$, the minimax Frobenius loss over the matrix set $\mathbb{V}_{d_1,d_2}^{R_0}$ is lower bounded as*

$$\mathcal{M}(\mathbb{V}_{d_1,d_2}^{R_0}) \ge c_1 R_0 \frac{\max\{d_1, d_2\}}{n}. \tag{10}$$

Proof. The proof of this lower bound follows standard procedures in information-theoretic analysis. Specifically, we first reduce the estimation problem on the minimax rate to a multiway hypothesis testing problem over a suitable packing set, and then Fano's inequality [8] is applied to lower bound the error probability. Define the Stiefel manifold $\mathbb{V}_{d_1,d_2}$ as follows

$$\mathbb{V}_{d_1,d_2} := \{\Theta \in \mathbb{R}^{d_1\times d_2} | \Theta^\top\Theta = \mathbb{I}_{d_2}\}. \tag{11}$$

Packing sets for the Stiefel manifold $\mathbb{V}_{d_1,d_2}$ will be utilized to construct the suitable packing set for the target set $\mathbb{V}_{d_1,d_2}^{R_0}$ (cf. (4)).

For a positive number $\Delta > 0$, let $M_F(\Delta)$ denote the cardinality of a maximal Δ-packing set contained in $\mathbb{V}_{d_1,d_2}^{R_0}$ in the Frobenius norm with elements $\{\Theta^1, \Theta^2, \cdots, \Theta^{M_F(\Delta)}\}$, and M is used as shorthand for $M_F(\Delta)$ in the following. It follows immediately from the standard technique in [8] that estimation on lower bound can be transformed into a multi-way hypothesis testing problem as

$$\mathbb{P}\left(\mathcal{M}(\mathbb{V}_{d_1,d_2}^{R_0}) \ge \frac{1}{4}\Delta^2\right) \ge \min_{\tilde{\Theta}}\mathbb{P}(\tilde{\Theta} \ne B), \tag{12}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 October 2022 doi:10.20944/preprints202210.0169.v1

where $\tilde{\Theta}$ is an estimator taking values in the packing set $\{\Theta^1, \Theta^2, \cdots, \Theta^M\}$, and $B \in \mathbb{R}^{d_1 \times d_2}$ is uniformly distributed at random over the packing set. Then one has by Fano's inequality [8] that

$$\mathbb{P}(\tilde{\Theta} \neq B) \geq 1 - \frac{I(Y;B) + \log 2}{\log M}, \tag{13}$$

where $I(Y;B)$ is the mutual information between the random distributed matrix B and the observation matrix $Y \in \mathbb{R}^{n \times d_2}$. It now remains to upper bound the mutual information $I(Y;B)$. Denote $\mathbb{P}_{\Theta^j}$ as the distribution of Y given $B = \Theta^j$ when Z is observed. Noting that Y has the mixture distribution $\frac{1}{M}\sum_{j=1}^{M} \mathbb{P}_{\Theta^j}$, we obtain that

$$\begin{aligned} I(Y;B) = \mathbb{E}_B\Big[D(\mathbb{P}_{Y|B}||\mathbb{P}_Y)\Big] &= \frac{1}{M}\sum_{j=1}^{M} D\left(\mathbb{P}_{\Theta^j}\middle|\middle|\frac{1}{M}\sum_{k=1}^{M}\mathbb{P}_{\Theta^k}\right) \\ &\leq \frac{1}{M^2}\sum_{j,k=1}^{M} D(\mathbb{P}_{\Theta^j}||\mathbb{P}_{\Theta^k}), \end{aligned} \tag{14}$$

where the last inequality is due to the convexity of the KL divergence. In what follows, we shall find a suitable packing for the set $\mathbb{V}_{d_1,d_2}^{R_0}$ to upper bound (14) by virtue of Lemma 2, and then to ensure that (13) is strictly larger than 0. Since $2 \leq R_0 \leq d_1 - R_0$ by assumption, [28, Lemma A.6 and Proposition 2.2] are applicable to concluding that for a positive number $\delta > 0$, there exists a subset $\{\tilde{\Theta}^1, \tilde{\Theta}^2, \cdots, \tilde{\Theta}^M\} \subseteq \mathbb{V}_{d_1,R_0}$ such that $\left|\left|\left|\tilde{\Theta}^j - \tilde{\Theta}^k\right|\right|\right|_F \geq \sqrt{R_0}\delta$ for all $j \neq k$, and $\log M \geq R_0(d_1 - R_0)\log(c_2/\delta)$, where c_2 is a universal positive constant. For $j = 1, \cdots, M$, set

$$\Theta^j = \begin{pmatrix} \tilde{\Theta}^j & 0 \end{pmatrix}, \tag{15}$$

where 0 stands for the $d_1 \times (d_2 - R_0)$ zero submatrix. Then it is easy to check that $\{\Theta^1, \Theta^2, \cdots, \Theta^M\} \subseteq \mathbb{V}_{d_1,d_2}^{R_0}$ and is a $\sqrt{R_0}\delta$-packing for $\mathbb{V}_{d_1,d_2}^{R_0}$ since $\left|\left|\left|\Theta^j - \Theta^k\right|\right|\right|_F = \left|\left|\left|\tilde{\Theta}^j - \tilde{\Theta}^k\right|\right|\right|_F \geq \sqrt{R_0}\delta$ for all $j \neq k$. Moreover, one has that $|||\Theta^j|||_F = |||\tilde{\Theta}^j|||_F = \sqrt{R_0}$, and thus $\left|\left|\left|\Theta^j - \Theta^k\right|\right|\right|_F \leq 2\sqrt{R_0}$. Set $\Delta = \sqrt{R_0}\delta$, then $\{\Theta^1, \Theta^2, \cdots, \Theta^M\}$ defined by (15) is just the Δ-packing of $\mathbb{V}_{d_1,d_2}^{R_0}$ that we are looking for, and the specific value of Δ will be determined later. Combining these arguments with Lemma 2, one sees that there exist universal positive constants (c_0, c_3) such that, (14) is upper bounded as

$$I(Y;B) \leq \frac{c_3 n \sigma_x^4}{\sigma_z^2 \sigma_\epsilon^2}(R_0 + d_2), \tag{16}$$

with probability at least $1 - 2\exp(-c_0 n + \log d_2)$. Define the random event $\mathscr{A}$={(16) happens}. Then it holds that $\mathbb{P}(\mathscr{A}) \geq 1 - 2\exp(-c_0 n + \log d_2)$. Substituting (16) into (13) yields that

$$\mathbb{P}(\tilde{\Theta} \neq B|\mathscr{A}) \geq 1 - \frac{\frac{c_3 n \sigma_x^4}{\sigma_z^2 \sigma_\epsilon^2}(R_0 + d_2) + \log 2}{R_0(d_1 - R_0)\log(c_0/\delta)}. \tag{17}$$

Set $\delta = 2c_4\sqrt{\frac{\max\{d_1,d_2\}}{n}}$ for a universal positive constant c_4, and thus $\Delta = 2c_4\sqrt{R_0\frac{\max\{d_1,d_2\}}{n}}$. Then if appropriate constants are chosen, (17) is strictly above zero, say bounded below by $1/2$. Specifically, it is easy to see that as long as the constants (c_2, c_3) are chosen to satisfy

$$\frac{c_3 n \sigma_x^4}{\sigma_z^2 \sigma_\epsilon^2}(R_0 + d_2) \leq \frac{1}{4}R_0(d_1 - R_0)\log(c_2/\delta), \text{ and} \tag{18a}$$

$$\log 2 \leq \frac{1}{4}R_0(d_1 - R_0)\log(c_2/\delta), \tag{18b}$$

then (17) is lower bounded by $1/2$. Indeed, since $2 \leq R_0 \leq d_1 - R_0$, one has that $R_0(d_1 - R_0) \geq 4$, and (18b) holds if $c_2 \geq 2\delta$. Then (18a) follows providing that $c_3 \leq$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 October 2022 doi:10.20944/preprints202210.0169.v1

$\frac{\sigma_z^2\sigma_\epsilon^2}{n\sigma_x^4(R_0+d_2)}\log 2$. Combining these arguments with (12), we obtain that there exist universal positive constants (c_0, c_1) such that

$$\mathbb{P}\left(\mathcal{M}(\mathbb{V}_{d_1,d_2}^{R_0}) \geq c_1 R_0 \frac{\max\{d_1,d_2\}}{n}\right) \geq \min_{\tilde{\Theta}} \mathbb{P}(\tilde{\Theta} \neq B) \geq \min_{\tilde{\Theta}} \mathbb{P}(\tilde{\Theta} \neq B|\mathscr{A})\mathbb{P}(\mathscr{A})$$
$$\geq \frac{1}{2}(1 - 2\exp(-c_0 n + \log d_2)).$$

The proof is complete. □

Remark 1. (i) *Theorem 1 tells us that in the additive noise case, with high probability, about* $\max\{d_1,d_2\}R_0$ *number of samples are required to estimate a* $d_1 \times d_2$ *matrix of rank* R_0 *consistently by any method. Note that the lower bound (10) agrees with the upper bound obtained in our another work [25, Theorem 1] when* $\lambda = \Omega\left(\sqrt{\frac{\max\{d_1,d_2\}}{n}}\right)$ *up to constant factors, implying that the proposed error-corrected estimator in [25] is minimax optimal in the additive noise case.*

(ii) *Researchers have investigated the matrix completion problem and established the information-theoretic limitations [12,14]. Specifically, for a* $d \times d$ *square matrix with rank* R_0*, Candès and Tao [12] showed that the samples needed to recover this matrix is of the order* $R_0 d \log d$*, while in [14], this order turns to* $R_0 d$ *with the additional "spikiness" imposed which refers to certain conditions on singular vectors of the low-rank matrix. Ja and Kutzarova [15] utilized the Gelfand widths of certain identity mappings between finite-dimensional Schatten p-spaces and showed the worst-case error for low-rank matrix recovery scales as* $R_0 d/n$*. Our minimax result is applicable to the more general multi-response regression model without the square matrix assumption on the underlying parameter, and still recaptures the above minimax rate. The established lower bound furthermore indicates that though in the additive noise case, no more samples are needed so as to achieve a rate-optimal estimation.*

4. Conclusion

We focused on the information-theoretic limitations of low-rank estimation for multi-response errors-in-variables regression under the high-dimensional scaling. The lower bound for the squared Frobenius loss over a special set of low-rank matrices was established by virtue of the information theory and statistical techniques. This lower bound matches the upper bound derived in our another work [18] considering the statistical estimation up to constant factors. Further research may generalize the current result to some more general classes or covariates that are sub-Gaussian with non-diagonal covariances. In addition, other types of measurement errors can also be considered, such as the multiplicative or dependent errors. The key to solving all these problems relies on a more delicate estimation of the KL divergence under different model assumptions.

Author Contributions: Conceptualization, Xin Li and Dongya Wu; Formal analysis, Xin Li; Funding acquisition, Xin Li and Dongya Wu; Investigation, Xin Li and Dongya Wu; Methodology, Xin Li; Project administration, Xin Li; Supervision, Dongya Wu; Validation, Xin Li and Dongya Wu; Writing – original draft, Xin Li; Writing – review & editing, Xin Li and Dongya Wu. All authors have read and agreed to the published version of the manuscript.

Funding: The APC was funded by the National Natural Science Foundation of China (12201496, 62103329), and the Natural Science Foundation of Shaanxi Province of China (2022JQ-045).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

 doi:10.20944/preprints202210.0169.v1

References

1. Bühlmann, P.; Van De Geer, S. *Statistics for high-dimensional data: Methods, theory and applications*; Springer Science & Business Media, 2011.
2. Wainwright, M.J. *High-dimensional statistics: A non-asymptotic viewpoint*; Vol. 48, Cambridge University Press, 2019.
3. Negahban, S.; Wainwright, M.J. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *Ann. Stat.* **2011**, pp. 1069–1097.
4. Zhou, H.; Li, L.X. Regularized matrix regression. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **2014**, *76*, 463–483.
5. Wu, D.Y.; Li, X.; Feng, J. Connectome-based individual prediction of cognitive behaviors via graph propagation network reveals directed brain network topology. *J. Neural Eng.* **2021**, *18*.
6. Wu, D.Y.; Li, X.; Feng, J. Multi-hops functional connectivity improves individual prediction of fusiform face activation via a graph neural network. *Front. Neurosci.* **2021**, *14*.
7. Loh, P.L. On lower bounds for statistical learning theory. *Entropy* **2017**, *19*, 617.
8. Yang, Y.H.; Barron, A. Information-theoretic determination of minimax rates of convergence. *Ann. Stat.* **1999**, *27*, 1564–1599.
9. Ye, F.; Zhang, C.H. Rate minimaxity of the Lasso and Dantzig selector for the ℓ_q loss in ℓ_r balls. *J. Mach. Learn. Res.* **2010**, *11*, 3519–3540.
10. Raskutti, G.; Wainwright, M.J.; Yu, B. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Trans. Inform. Theory* **2011**, *57*, 6976–6994.
11. Wang, Z.; Paterlini, S.; Gao, F.C.; Yang, Y.H. Adaptive minimax regression estimation over sparse ℓ_q-hulls. *J. Mach. Learn. Res.* **2014**, *15*, 1675–1711.
12. Candès, E.J.; Tao, T. The power of convex relaxation: Near-optimal matrix completion. *IEEE Trans. Inform. Theory* **2010**, *56*, 2053–2080.
13. Koltchinskii, V.; Tsybakov, A.B.; Lounici, K. Nuclear norm penalization and optimal rates for noisy low rank matrix completion. *Ann. Stat.* **2010**, *39*.
14. Negahban, S.; Wainwright, M.J. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *J. Mach. Learn. Res.* **2012**, *13*, 1665–1697.
15. Ja, C.D.; Kutzarova, D. Stability of low-rank matrix recovery and its connections to Banach space geometry. *J. Math. Anal. Appl.* **2015**, *427*, 320–335.
16. Sørensen, Ø.; Frigessi, A.; Thoresen, M. Measurement error in Lasso: Impact and likelihood bias correction. *Stat. Sci.* **2015**, *25*, 809–829.
17. Datta, A.; Zou, H. Cocolasso for high-dimensional error-in-variables regression. *Ann. Stat.* **2017**, *45*, 2400–2426.
18. Li, X.; Wu, D.Y.; Li, C.; Wang, J.H.; Yao, J.C. Sparse recovery via nonconvex regularized M-estimators over ℓ_q-balls. *Comput. Stat. Data Anal.* **2020**, *152*, 107047.
19. Loh, P.L.; Wainwright, M.J. High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *Ann. Stat.* **2012**, *40*, 1637–1664.
20. Rosenbaum, M.; Tsybakov, A.B. Sparse recovery under matrix uncertainty. *Ann. Stat.* **2010**, *38*, 2620–2651.
21. Loh, P.L.; Wainwright, M.J. Corrupted and missing predictors: Minimax bounds for high-dimensional linear regression. IEEE International Symposium on Information Theory Proceedings, 2012, pp. 2601–2605.
22. Li, X.; Wu, D.Y. Minimax rates of ℓ_p-losses for high-dimensional linear errors-in-variables models over ℓ_q-balls. *Entropy* **2021**, *23*, 722.
23. Recht, B.; Fazel, M.; Parrilo, P.A. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Rev.* **2010**, *52*, 471–501.
24. Antonacci, Y.; Astolfi, L.; Nollo, G.; Faes, L. Information transfer in linear multivariate processes assessed through penalized regression techniques: Validation and application to physiological networks. *Entropy* **2020**.
25. Li, X.; Wu, D.Y. Low-rank matrix estimation in multi-response regression with measurement errors: Statistical and computational guarantees. *arXiv preprint arXiv:2012.05432v3* **2020**.
26. Carroll, R.J.; Ruppert, D.; Stefanski, L.A.; Crainiceanu, C.M. *Measurement error in nonlinear models: A modern perspective, second ed*; Chapman & Hall/CRC: Boca Raton, Florida, 2006.
27. Loh, P.L.; Wainwright, M.J. Supplementary material: High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *Ann. Stat.* **2012**.
28. Vu, V.Q.; Lei, J. Minimax sparse principal subspace estimation in high dimensions. *Ann. Stat.* **2013**, *41*, 2905–2947.