Statistics and Computing (2024) 34:10
https://doi.org/10.1007/s11222-023-10318-z

ORIGINAL PAPER

A framework of regularized low-rank matrix models for regression and classification

Hsin-Hsiung Huang[1] · Feng Yu[2] · Xing Fan[3] · Teng Zhang[3]

Received: 18 July 2023 / Accepted: 27 September 2023

Abstract
While matrix-covariate regression models have been studied in many existing works, classical statistical and computational methods for the analysis of the regression coefficient estimation are highly affected by high dimensional matrix-valued covariates. To address these issues, this paper proposes a framework of matrix-covariate regression models based on a low-rank constraint and an additional regularization term for structured signals, with considerations of models of both continuous and binary responses. We propose an efficient Riemannian-steepest-descent algorithm for regression coefficient estimation. We prove that the consistency of the proposed estimator is in the order of $O(\sqrt{r(q+m)+p}/\sqrt{n})$, where r is the rank, $p \times m$ is the dimension of the coefficient matrix and p is the dimension of the coefficient vector. When the rank r is small, this rate improves over $O(\sqrt{qm+p}/\sqrt{n})$, the consistency of the existing work (Li et al. in Electron J Stat 15:1909-1950, 2021) that does not apply a rank constraint. In addition, we prove that all accumulation points of the iterates have similar estimation errors asymptotically and substantially attaining the minimax rate. We validate the proposed method through a simulated dataset on two-dimensional shape images and two real datasets of brain signals and microscopic leucorrhea images.

Keywords Electroencephalography (EEG) · Generalized linear model (GLM) · High dimensionality · ℓ_1 norm · Microscopic leucorrhea images · Rank constraint · Riemannian steepest descent · Sparsity

1 Introduction

In numerous modern scientific applications, matrix-valued covariates of interest naturally exist in massive datasets, for example, a grayscale image that quantifies the intensities of image pixels is represented as a two-dimensional (2D) data matrix (Zhou and Li 2014), and the recommendation system of online service records the preferences of users over their products as a matrix (Elsener and Geer 2018). Other applications include the brain signal electroencephalography (EEG) dataset (Hung and Wang 2012; Zhou and Li 2014), the microscopic leucorrhea images (Hao et al. 2019), and the diabetes data (Li et al. 2021). To analyze such datasets, it is of great interest to investigate matrix-covariate regression and logistic regression models with regularization. While we may vectorize a matrix-valued covariate as a vector and apply standard procedures, the size of the vector might be large and standard algorithms are inefficient and computationally expensive.

To induce such a sparse or low-rank structure, a variety of regularization-based matrix regression tools have been recently proposed. For example, nuclear-norm penalized optimization has been applied in several works (Elsener and Geer 2018; Lu et al. 2012; Negahban and Wainwright 2011; Fan et al. 2021) to induce a low-rank structure. In particular, (Negahban and Wainwright 2011) considered a standard M-estimator based on regularization by the nuclear or trace norm over matrices and analyzed its performance under a high-dimensional setting, and (Elsener and Geer 2018) con-

✉ Teng Zhang
teng.zhang@ucf.edu

Hsin-Hsiung Huang
hsin-hsiung.huang@ucf.edu

Feng Yu
fyu@umn.edu

Xing Fan
fanxing@knights.ucf.edu

[1] Department of Statistics and Data Science, University of Central Florida, 4000 Central Florida Blvd, Orlando, FL 32816, USA

[2] School of Mathematics, University of Minnesota Twin Cities, Vincent Hall 206 Church St., Minneapolis, MN 55455, USA

[3] Department of Mathematics, University of Central Florida, 4000 Central Florida Blvd, Orlando, FL 32816, USA

Published online: 25 October 2023

sidered robust nuclear norm penalized estimators using two well-known robust loss functions: the absolute value loss and the Huber loss, and derive the asymptotic performance of these estimators. (Fan et al. 2021) applied the nuclear-norm penalized least-squares approach to appropriately truncated or shrunk data to four popular problems: sparse linear models (Tibshirani 1996), compressed sensing (Donoho 2006), matrix completion (Candès and Recht 2009), and multitask learning (Caruana 1997), as well as robust covariance estimation (Campbell 1980).

Our work has three main contributions. First, we investigate a framework of matrix regression problems with a low-rank constraint and ℓ_1 or total variation (TV) regularization. Second, we propose a computationally efficient Riemannian gradient descent algorithm that has a smaller or comparable computational cost than existing methods and has a convergence guarantee. Third, we establish theoretical guarantees by showing that the proposed estimates are consistent. In addition, we show that when the rank is small, our estimation error $O(\sqrt{r(q+m)+p}/\sqrt{n})$ is smaller than $O(\sqrt{qm+p}/\sqrt{n})$, the estimation error by regularization-based methods without rank constraints (Li et al. 2021), and acquires the desired statistical accuracy in a minimax sense (Tsybakov 2008; Koltchinskii et al. 2011; She et al. 2021). Our consistency theorem (Theorem 4) does not depend on the convexity of the loss function, and hence it applies to various choices of loss functions, including robust loss functions such as redescending ψ's (Huber 1964), Hampel's loss, or Tukey's bisquare, etc. (Maronna et al. 2018; She and Chen 2017; Huang and Zhang 2020).

2 Methodology

Following existing works such as Zhou and Li (2014), Rohde and Tsybakov (2011) and Li et al. (2021), we use a matrix-covariate regression model that includes a coefficient matrix $\mathbf{C}^* \in \mathbb{R}^{m\times q}$ and a coefficient vector $\gamma^* \in \mathbb{R}^p$. We remark that the trace regression model in Rohde and Tsybakov (2011) and Elsener and Geer (2018) can be considered as a special case of this model when γ^* is zero.

In regression problems, the responses are continuous univariate variables. For such applications, we assume that for each $1 \le i \le n$, the i-th response y_i is generated from the matrix predictor $\mathbf{X}_i \in \mathbb{R}^{m\times q}$ and $\mathbf{z}_i \in \mathbb{R}^p$ by

$$y_i = \langle \mathbf{X}_i, \mathbf{C}^* \rangle + \langle \mathbf{z}_i, \gamma^* \rangle + \epsilon_i, \tag{2.1}$$

where $\langle \mathbf{X}_i, \mathbf{C}^* \rangle$ is defined as the trace of $\mathbf{C}^{*T}\mathbf{X}_i$, $\langle \mathbf{z}_i, \gamma^* \rangle$ is defined as $\mathbf{z}_i^T \gamma^*$, and $\{\epsilon_i\}_{i=1}^n$ are the observation errors.

In binary classification problems, the responses are binary with value 1 or 0 to indicate the presence or absence of the target category. For such applications, we follow the matrix-covariate logistic regression model from Hung and Wang (2012) and Li et al. (2021) and assume that the response y_i is a binary variable such that

$$\begin{aligned} &\text{logit}\left(\Pr(y_i = 1 \mid \mathbf{X}_i)\right) \\ &= \log\left(\frac{\Pr(y_i = 1|\mathbf{X}_i)}{1 - \Pr(y_i = 1|\mathbf{X}_i)}\right) = \langle \mathbf{X}_i, \mathbf{C}^* \rangle + \langle \mathbf{z}_i, \gamma^* \rangle, \end{aligned} \tag{2.2}$$

and the explicit formula is

$$\begin{aligned} &y_i \sim \text{Bernoulli}(p_i), \text{ where } p_i = \frac{e^{\theta_i}}{1+e^{\theta_i}} \text{ and} \\ &\theta_i = \langle \mathbf{X}_i, \mathbf{C}^* \rangle + \langle \mathbf{z}_i, \gamma^* \rangle. \end{aligned} \tag{2.3}$$

The matrix-covariate regression model (2.1) and the matrix-covariate logistic regression model (2.2) are generalizations of the regular regression model and the regular logistic regression model. We aim to estimate the coefficient matrix $\mathbf{C}^* \in \mathbb{R}^{m\times q}$ and the coefficient vector $\gamma^* \in \mathbb{R}^p$, based on covariates $\{\mathbf{X}_i, \mathbf{z}_i\}_{i=1}^n \subset \mathbb{R}^{m\times q} \times \mathbb{R}^p$ and the associated responses $\{y_i\}_{i=1}^n$ generated from the matrix-covariate regression model (2.1) or the matrix-covariate logistic regression model (2.2).

In this work, $\mathbf{C}^*$ is assumed to have a low-rank structure (or can be approximated by a low-rank structure), and we propose an estimator under the rank constraint as follows:

$$(\hat{\mathbf{C}}, \hat{\gamma}) = \underset{\gamma\in\mathbb{R}^p, \mathbf{C}\in\mathbb{R}^{m\times q}, \text{rank}(\mathbf{C})=r}{\text{argmin}} F(\mathbf{C}, \gamma), \tag{2.4}$$

where $F(\mathbf{C}, \gamma) = \sum_{i=1}^n l(y_i, \langle \mathbf{X}_i, \mathbf{C} \rangle + \gamma^T \mathbf{z}_i) + \lambda P(\mathbf{C}, \gamma)$, $l(\cdot,\cdot)$ is a loss function that measures the difference between the observed response y_i and predicted response $\langle \mathbf{X}_i, \mathbf{C} \rangle + \gamma^T \mathbf{z}_i$, and $P(\cdot)$ is a penalty function for $(\mathbf{C}, \gamma)$. For the matrix-covariate regression model (2.1) and the binary response model following Eq. (2.2), we apply different loss functions as follows.

- *The matrix-covariate regression model with continuous responses*. For model (2.1), we apply the least-square loss function $l(y_i, f_i) = \frac{1}{2}(y_i - f_i)^2$. We remark that we can also apply a robust loss function such as Huber's loss function (Huber 1964).
- *The binary response and matrix-covariate logistic regression model*. For model (2.2), we apply the logistic loss function

$$l(y_i, f_i) = \log(1 + \exp(f_i)) - y_i f_i.$$

In addition, the penalty function P in $F(\mathbf{C}, \gamma)$ of Eq. (2.4) can be chosen according to specific applications and the prior knowledge of $\mathbf{C}^*$ and γ^*. For example, if $\mathbf{C}^*$ is known to be

sparse element-wisely, we may apply $P(\mathbf{C}, \gamma) = \|\mathbf{C}\|_1$. If $\mathbf{C}^*$ is known to be piecewise constant, we may apply the total variation penalty $P(\mathbf{C}, \gamma) = \|\mathbf{C}\|_{TV} = \sum_{i,j} |\mathbf{C}_{i,j} - \mathbf{C}_{i+1,j}| + |\mathbf{C}_{i,j} - \mathbf{C}_{i,j+1}|$. In this work, we leave the choice open for the theoretical analysis and choose the penalty P accordingly in experiments.

2.1 An efficient parameter estimation algorithm

The minimization problem (2.4) can be considered as a manifold-constrained optimization in the form of

$$\underset{\mathbf{C},\gamma}{\operatorname{argmin}} F(\mathbf{C}, \gamma) \quad \text{s.t. } (\mathbf{C}, \gamma) \in \mathcal{M} \times \mathbb{R}^p, \tag{2.5}$$

where $\mathcal{M} = \{\mathbf{C} \in \mathbb{R}^{m \times q} \mid \operatorname{rank}(\mathbf{C}) = r\}$ is the fixed-rank manifold. Since it is defined as a product of two manifolds, $\mathcal{M} \times \mathbb{R}^p$, Riemannian approaches can be applied to solving (2.4), and we refer readers (Absil et al. 2009) and (Boumal et al. 2014a) for more technical details. Here we apply the Riemannian steepest descent method and our implementation is based on package `manopt` (Boumal et al. 2014b). The algorithm is described in Algorithm 1. It depends on the projection $\mathcal{R}$ at $\mathbf{C}$, denoted by $\mathcal{R}_{\mathbf{C}}$, which is a mapping from $\mathbb{R}^{m \times q}$ to $\mathcal{M}$ that preserves local gradients at $\mathbf{C}$, and has an explicit expression (Vandereycken 2013).

For Algorithm 1, we establish a theoretical guarantee that it converges to a stationary point of the objective function as follows. Here $\|\cdot\|_F$ is the Frobenius norm.

Algorithm 1 Regularized low-rank matrix-covariate regression

Input: The samples $\{(y_i, \mathbf{X}_i, \mathbf{z}_i)\}_{i=1}^n$, initial $(\mathbf{C}^{(0)}, \gamma^{(0)})$. Error tolerance ϵ and maximum iteration K.
1: For $k = 0, 1, 2, \cdots$ do
2: Set $(\zeta_1, \zeta_2) = (\partial_{\mathbf{C}} F(\mathbf{C}^{(k)}, \gamma^{(k)}), \partial_{\gamma} F(\mathbf{C}^{(k)}, \gamma^{(k)}))$.
3: Use line search to select a step size α^k such that $F(\mathcal{R}_{\mathbf{C}^{(k)}}(\alpha^k \zeta_1), \gamma(k) - \alpha^k \zeta_2) \le F(\mathbf{C}^{(k)}, \gamma^{(k)})$.
4: Set $(\mathbf{C}^{(k+1)}, \gamma^{(k+1)}) = (\mathcal{R}_{\mathbf{C}^{(k)}}(\alpha^k \zeta_1), \gamma^{(k)} - \alpha^k \zeta_2)$.
5: Stop until $(F(\mathbf{C}^{(k)}, \gamma^{(k)}) - F(\mathbf{C}^{(k+1)}, \gamma^{(k+1)}))/F(\mathbf{C}^{(k)}, \gamma^{(k)}) \le \epsilon$ or $k \ge K$
Output: Set $\hat{\mathbf{C}} = \mathbf{C}^{(k)}, \hat{\gamma} = \gamma^{(k)}$.

Proposition 1 *(Algorithmic convergence) (a) The functional values $F(\mathbf{C}^{(\text{iter})}, \gamma^{(\text{iter})})$ are nonincreasing and converge. (b) Each accumulation point of the sequence $(\mathbf{C}^{(\text{iter})}, \gamma^{(\text{iter})})$ is a point where the subgradients of the objective function $F(\mathbf{C}, \gamma)$ contain zero, and is a stationary point of F when F is smooth.*

Algorithm 1 can be implemented efficiently by storing $\mathbf{C}^{(k)}$ using its low-rank expression. Then both step 2 and step 3 require computational costs per iteration in the order of $O(nr(q+m) + np + qm)$, and step 4 requires $O(rqm)$. Combining them, Algorithm 1 has a computational cost per iteration of $O(nr(q+m) + np + qm)$, which is an improvement over the computational cost of the existing algorithm without a rank constraint (Li et al. 2021) when r is small, which is given by $O(nqm + np)$.

3 Asymptotic theory

This section is devoted to analyzing the statistical consistency property of the estimator (2.4) and Algorithm 1. First, Sect. 3.1 presents the consistency rate of the proposed estimator (2.4) in Theorem 2. Second, Sect. 3.2 presents our result on the minimax rate of the estimation problem (2.4) in Theorem 3. Finally, Sect. 3.3 proves the statistical convergence rate of Algorithm 1 when the algorithm is well-initialized in Theorem 4.

3.1 Consistency of the proposed estimator

This section establishes the consistency of the proposed estimator (2.4) in Theorem 2 and shows that for large n, our estimator converges to the true underlying solution $(\mathbf{C}^*, \gamma^*)$. This result holds for general penalty P and regularization parameter λ. The detailed proof is deferred to the appendix. This consistency result depends on Assumptions 1–3 that are described in the appendix in detail, where Assumptions 1–2 can be considered as a generalization of the commonly used restricted isometry property (RIP) (Candes and Tao 2005; Recht et al. 2010) to the setting of matrix regression. We remark that these assumptions ensure that the landscape of the objective function is well-behaved around the true solution $(\mathbf{C}^*, \gamma^*)$, and similar assumptions also appeared in the literature of matrix regression in (Li et al. 2021, Conditions 1,2,5,6).

Theorem 2 *(Consistency of the proposed estimator) Under Assumptions 1–3 (see the appendix), then for some large constant C and assume that*

$$n \ge C\left(r(qm+p) + \lambda P(\mathbf{C}^*, \gamma^*)\right) \tag{3.1}$$

and for all $t \ge 2$, we have the following upper bound on the estimation error of the proposed estimator (2.4) with probability at least $1 - C\exp(-Cn) - C\exp(-Ct(r(q+m)+p))$:

$$\operatorname{dist}((\hat{\mathbf{C}}, \hat{\gamma}), (\mathbf{C}^*, \gamma^*)) \le Ct\sqrt{\frac{(r(q+m)+p)}{n}} + C\sqrt{\frac{\lambda P(\mathbf{C}^*, \gamma^*)}{n}}. \tag{3.2}$$

Here $\operatorname{dist}(\cdot,\cdot)$ is the distance induced by the Frobenius norm.

Comparison of the convergence rate with existing works. The main result, Inequality (3.2), shows that the estimation error is in the order of

$$O_P\left(\sqrt{\frac{r(q+m)+p}{n}}+\sqrt{\frac{\lambda P(\mathbf{C}^*,\gamma^*)}{n}}\right).$$

When λ is bounded by a constant, then the first term is the dominant term and the estimation error is in the order of $O_P\left(\sqrt{\frac{r(q+m)+p}{n}}\right)$. In comparison, the rate in the existing work (Li et al. 2021) is $O_P\left(\sqrt{\frac{qm+p}{n}}+\sqrt{\frac{\lambda P(\mathbf{C}^*,\gamma^*)}{n}}\right)$. Our result improves the factor $mq+p$ to $r(q+m)+p$, which is a significant improvement when rank r is smaller than q and m. This improvement can be explained by the fact that by fixing the rank constraint, our estimator has $r(q+m)+p$ parameters which are fewer than the $qm+p$ parameters in the estimator of Li et al. (2021). With fewer parameters to estimate, our estimator achieves a better convergence rate, and this improvement is also clear from our simulation experiments. O_P and o_P represent the standard big and small O notation in probability.

Comparison with minimax rate. We derive our minimax analysis in Theorem 4, which shows that under the setting that $\mathbf{C}^*$ is low-rank, the minimax rate of estimating $(\mathbf{C},\gamma)$ is in the order of $O_P\left(\sqrt{\frac{r(q+m)+p}{n}}\right)$. In comparison, our rate in Theorem 2 achieves this rate when the regularization parameter λ is zero or bounded by $O(1)$. In literature, (Luo et al. 2020) shows that the minimax rate is in the order of $O_P\left(\sqrt{\frac{r(q+m)}{n}}\right)$ when the vector covariate γ does not exist and $p=0$ and can be considered as a special case of our result. In the future, we plan to derive an improved minimax rate of our model under the setting that $\mathbf{C}^*$ is both sparse and low-rank based on the technique of She et al. (2021), and show that our estimator with a well-chosen $\lambda\neq 0$ and $P(\mathbf{C},\gamma)=\|\mathbf{C}\|_1$ achieves the improved rate.

MLE argument when $\lambda=0$ In addition, the adaption of the usual arguments of MLE to manifold optimization can be applied when $\lambda=0$, in which we may show that $\sqrt{n}(\tilde{\mathbf{C}}-\mathbf{C},\tilde{\gamma}-\gamma)$ converges to a Gaussian-like distribution (the result is similar to Theorem 3.1 of (Hung and Wang 2012)), where the covariance/Fisher information matrix can be obtained following the techniques used in Le Cam (1990) and Boumal (2013). However, we skip the rigorous statement here as we use $\lambda>0$ in practice.

A generic choice of P In this work, we leave the choice of the penalty P open for theoretical analysis. However, there are natural choices of P in certain applications. For example, when we have the prior knowledge that $\mathbf{C}^*$ is sparse, we may choose $P(\mathbf{C},\gamma)=\|\mathbf{C}\|_1$ for variable selection.

Choice of rank r The choice of r has a large impact on the solution. In fact, the objective value of (2.4) is nonincreasing as r increases, since the set of matrices of rank r_1 contains the set of matrices of rank r_2 when $r_2>r_1$. We expect that a choice of r can be obtained using either cross-validation or the elbow method (Choi et al. 2017) that chooses a rank such that a larger rank doesn't lead to a much smaller objective value.

3.2 Minimax rate of the estimation problem

This section proves that the rates attained by our estimator are optimal using the minimax lower bound of our estimation problem.

Theorem 3 *(Minimax rate) Consider the class of parameters*

$$\mathcal{A}(r,a)=\{(\mathbf{C},\gamma)\in\mathbb{R}^{m\times q}\times\mathbb{R}^p:\ \operatorname{rank}(\mathbf{C})\le r,\ \|\mathbf{C}\|_F^2+\|\gamma\|^2\le a\}.$$

Under Assumptions 5–6 (see the appendix) and assume in addition that $n\ge r(m+q)+p$, then for any $\beta\in\left(0,\frac{2^{(r(m+q)+p)/4}}{1+2^{(r(m+q)+p)/4}}\right)$, there exists $c_0>0$ depending on β such that

$$\inf_{\hat{\mathbf{C}},\hat{\gamma}}\sup_{(\mathbf{C}^*,\gamma^*)\in\mathcal{A}(r,a)}\Pr\left(\operatorname{dist}\left((\mathbf{C}^*,\gamma^*),(\hat{\mathbf{C}},\hat{\gamma})\right)\ge c_0\sqrt{\frac{r(m+q)+p}{n}}\right)\ge\beta.$$

Combining it with the convergence rate in Theorem 2, it implies that the rate of our estimator in Theorem 2 is optimal when λ is reasonably chosen, such as zero or a parameter smaller than $r(q+m)+p$. The proof is based on the arguments used by Tsybakov (2008) and is deferred to the appendix.

3.3 Convergence of the proposed algorithm

This section shows that the output of Algorithm 1 is sufficiently close to the ground truth with a good initialization and an additional assumption on the second-order information of the "Hessian matrix".

Theorem 4 *(Statistical convergence rate of the proposed algorithm) Under Assumptions 1–4, the initialization $(\mathbf{C}^{(0)},\gamma^{(0)})$ is "good" in the sense that both its distance to $(\mathbf{C}^*,\gamma^*)$, $\operatorname{dist}((\mathbf{C}^{(0)},\gamma^{(0)}),(\mathbf{C}^*,\gamma^*))$, and the objective value $F(\mathbf{C}^{(0)},\gamma^{(0)})$ are bounded by a constant, and the number of samples n is large: $n>C(qm+p)$. Then for all $t\ge 2$, with a probability at least $1-C\exp(-Cn)-C\exp(-t(r(q+m)+p))$, all accumulation points of the iterates $\{(\mathbf{C}^{(\text{iter})},\gamma^{(\text{iter})})\}_{\text{iter}\ge 1}$,*

denoted by $(\tilde{\mathbf{C}}, \tilde{\gamma})$*, have small estimation errors:*

$$\text{dist}((\tilde{\mathbf{C}}, \tilde{\gamma}), (\mathbf{C}^*, \gamma^*)) \leq C\frac{t\sqrt{r(q+m)+p}}{\sqrt{n}} + C\frac{\lambda C_{partial}}{n}, \tag{3.3}$$

where $C_{partial}$ *is the upper bound of the magnitude of the derivative of the penalty function within a neighborhood of the ground truth:* $C_{partial} = \max_{\{\text{dist}((\mathbf{C},\gamma),(\mathbf{C}^*,\gamma^*))\leq c\}} \sqrt{\left\|\frac{\partial}{\partial \mathbf{C}} P(\mathbf{C},\gamma)\right\|_F^2 + \left\|\frac{\partial}{\partial \gamma} P(\mathbf{C},\gamma)\right\|^2}$.

Remark 1 Unlike Theorem 2, Theorem 4 does not depend on the convexity of the loss function l. As a result, it applies to various choices of l, including popular loss function functions in robust statistics such as redescending ψ's, Hampel's loss, or Tukey's bisquare, etc. (Huber 1964; Maronna et al. 2018; She and Chen 2017; Huang and Zhang 2020) that can detect outliers with moderate or high leverages.

Convergence rate The key result in this section, (3.3), shows that the algorithm achieves a similar estimation error as the result in the consistency result in Inequality (3.2), with the term $\sqrt{\frac{6\lambda P(\mathbf{C}^*,\gamma^*)}{C_2 n}}$ replaced with $\frac{4\lambda C_{partial}}{nC_2}$. For many common penalty functions, $C_{partial}$ is bounded. For example, for the ℓ_1 loss function that $P(\mathbf{C}, \gamma) = \|\mathbf{C}\|_1$ used in our simulations, we have $C_{partial} \leq \sqrt{qm}$.

Condition on initialization Theorem 4 makes two assumptions on the initialization. The first condition on the distance is straightforward, and the second condition on the initial objective value is satisfied when

$$\text{dist}((\mathbf{C}^{(0)}, \gamma^{(0)}), (\mathbf{C}^*, \gamma^*)) = o_P(1)$$

and $\lambda = o(1)$ as $n \to \infty$, which can be justified using Assumption 4.

Obtaining initialization that satisfies the conditions Theorem 4 requires an initialization that is within a neighborhood of the true solution of radius $O(1)$. While empirically we initialize $\mathbf{C}^{(0)}$ and $\gamma^{(0)}$ as a zero matrix and a zero vector individually, it is possible to obtain initialization with theoretical guarantees. For example, for the matrix-covariate regression with the continuous response setting, we may let the initialization to be the solution of the standard regression problem, i.e., the solution of $\text{argmin}_{\mathbf{C},\gamma} \sum_{i=1}^n (y_i - \langle \mathbf{X}_i, \mathbf{C}\rangle - \gamma^T \mathbf{z}_i)^2$, which has an initial estimation error in the order of $O_P\left(\sqrt{\frac{qm+p}{n}}\right)$. For the matrix-covariate logistic regression setting, we may solve $\text{argmin}_{\mathbf{C},\gamma} \sum_{i=1}^n l(y_i, \langle \mathbf{X}_i, \mathbf{C}\rangle + \gamma^T \mathbf{z}_i)$ as well. These are convex problems, and the standard asymptotic analysis shows that the estimation errors would converge to zero as $n \to \infty$, that is, the initialization conditions will be satisfied as $n \to \infty$.

Convergence rate Unfortunately, it is difficult to obtain a general result of the convergence rate to $(\hat{\mathbf{C}}, \hat{\gamma})$ without assumptions on the penalty P. However, our proof implies that Algorithm 1 converges linearly to a neighborhood around the optimal solution $(\mathbf{C}^*, \gamma^*)$, and empirically Algorithm 1 converges quickly in our simulations.

4 Simulations

In this section, we carry out several numerical studies to investigate the empirical performance of our proposed methods on synthetic data, brain signal electroencephalography (EEG) data and leucorrhea data. For the continuous response and the matrix-covariate regression model, we compare our estimator with the spectral regularized regression estimator (SRRE) proposed in Zhou and Li (2014) and the low-rank estimation-testing matrix regression estimator (LEME) proposed in Hung and Jou (2019). In addition to assessing the accuracy of matrix estimation and prediction, we also compare the computational efficiency of our estimator with SRRE in terms of implementation time. For the binary response and the matrix-covariate logistic regression model, we compare our estimator with LEME and the SDNCMV method described in Chen et al. (2021).

It should be noted that although Li et al. (2021) introduced the double fused Lasso regularized matrix regression (DFMR) and its logistic version (DFMLR), we do not include a comparison with these methods in our study. This is because the DFMR and DFMLR assume a specific structure for the matrix-covariate $\mathbf{C}^*$ and the vector-covariate γ^*, where the difference between successive rows of $\mathbf{C}^*$ is sparse and the vector-covariate γ^* is also sparse.

4.1 Simulation I: matrix-covariate regression

In this section, we investigate the continuous response and the matrix-covariate regression model (2.1). The predictors $\mathbf{X}_i \in \mathbb{R}^{64\times 64}$ and $\mathbf{z}_i \in \mathbb{R}^5$, as well as the observation errors $\epsilon_i \in \mathbb{R}$, are randomly sampled from a standard Gaussian distribution $N(0, 1)$. We set $\gamma^* = (1, 1, 1, 1, 1)^T$, and the matrix $\mathbf{C}^* \in [0, 1]^{64\times 64}$ represents a 64 by 64 image displayed in the first column of Fig. 1.

4.1.1 Sparse penalty

To validate the consistency analysis presented in Theorem 2, we utilize the ℓ_1 penalty $P(\mathbf{C}) = \|\mathbf{C}\|_1$ and explore different sample sizes: $n = 300, 500, 700, 1000$. The RMRE method utilizes the true ranks of the first three images (square, T-shape, cross) that exhibit simple shapes, whereas approximate ranks (5 for triangle and circle, and 10 for butterfly) are employed for the last three images (triangle, circle, but-

terfly) with more intricate shapes. The selection of λ is performed using the validation set approach across all three methods: RMRE, LEME, and SRRE. The estimation results, presented in Table 1, include the average root-mean-square error (RMSE) $\sqrt{\|\hat{\mathbf{C}} - \mathbf{C}^*\|_F^2/qm}$ and its standard deviation. The findings reveal that RMRE consistently outperforms SRRE and LEME in terms of estimation accuracy, except for the circle and butterfly shapes when $n \leq 500$, which can be attributed to the non-low-rank nature of $\mathbf{C}^*$ and the complexity of the shapes. Moreover, the estimation errors demonstrate a decreasing pattern on the order of $1/\sqrt{n}$, thus providing evidence to support the consistency analysis outlined in Sect. 3.

Furthermore, Fig. 1 illustrates the estimated matrix-covariate $\hat{\mathbf{C}}$ using RMRE and SRRE with λ determined through the validation set approach, as well as the unregularized RMRE with $\lambda = 0$, for $n = 500$. Several observations can be made from Fig. 1: (a) RMRE with the low-rank constraint leads to improved estimation accuracy compared to SRRE, particularly for shapes such as square, T, and cross, where the true $\mathbf{C}^*$ is strictly low-rank. Notably, even the unregularized RMRE (with $\lambda = 0$) outperforms SRRE in estimating the matrix-covariate $\mathbf{C}^*$; (b) The regularization in RMRE further enhances the estimation quality for shapes such as triangle and circle compared to the unregularized RMRE; (c) Given the intricate nature of the butterfly shape, both RMRE and SRRE fail to produce clear images at $n = 500$, indicating the need for larger sample sizes to achieve more accurate estimations of $\mathbf{C}^*$ (as demonstrated in Table 1).

We conducted a comparison of the three methods in terms of computational time, and the results are depicted in Fig. 3. To ensure a fair assessment, we performed four sets of simulations using two different images (a square and a T-shape) that were resized to dimensions of $p \times p$. We varied the sample sizes for each simulation scenario. For comparison, we fine-tuned the parameters of the three methods using an independent dataset and measured the CPU times. The rank used in the RMRE is the true rank of the images. The curves in the figure represent the average results obtained from 30 repeated simulations. Notably, our proposed method, RMRE, along with LEME, exhibited significantly faster execution times compared to SRRE. However, RMRE has much smaller errors than LEME (see Table 1).

4.1.2 Total variation penalty

It is important to note that our method allows for various choices of regularization.

In addition to the ℓ_1-regularization utilized in Table 1 and Fig. 1, another natural option for image denoising is total variation (TV) regularization. We select the ℓ_1-based isotropic TV regularization Beck and Teboulle (2009), defined as $P(\mathbf{C}) = \sum_{i,j} |\mathbf{C}_{i,j+1} - \mathbf{C}_{i,j}| + |\mathbf{C}_{i+1,j} - \mathbf{C}_{i,j}|$.

To assess the performance of this variant of our estimator, we compare it with SRRE and SIG-TV, a generalized scalar-on-image regression model proposed by Wang et al. (2017) that incorporates TV regularization while considering $\mathbf{X}$ as an image predictor to predict a scalar response y. We do not include a comparison with LEME because although it enforces a low-rank structure on the signal, it does not take into account sparsity or small total variation. The predictors $\mathbf{X}_i \in \mathbb{R}^{64\times 64}$, $\mathbf{z}_i \in \mathbb{R}^5$, and the observation errors $\epsilon_i \in \mathbb{R}$ are randomly sampled from a standard Gaussian distribution $N(0, 1)$. The true signal $\mathbf{C}^*$ consists of four 2D images, and the number of samples used in this simulation varies with $n = 300, 350, 350, 1000$. We utilize the true ranks for the first three images (square, T-shape, and cross) and set the rank for the fourth image (Phantom image) as 20. Figure 2 displays the images recovered by our proposed method (RMRE) using the ℓ_1-regularization and TV-regularization, SIG-TV, SRRE, as well as the regression without any regularization. The results presented in Fig. 2 demonstrate that the TV-regularization produces similar effects to the ℓ_1-regularization in our model. For simpler images such as the T-shape, the low-rank constraint effectively regularizes the estimation and leads to improved accuracy. Moreover, these estimated images indicate that both the ℓ_1 and TV regularizations perform better than the nuclear norm regularization employed in SRRE.

4.2 Simulation II: matrix-covariate logistic regression

In this subsection, we consider the binary response and logistic regression model in Eq. (2.3) using simulation data. Similar to Sect. 4.1, the predictors $\mathbf{X}_i \in \mathbb{R}^{64\times 64}$ and $\mathbf{z}_i \in \mathbb{R}^5$, as well as the observation errors $\epsilon_i \in \mathbb{R}$, are elementwise sampled from the standard Gaussian distribution $N(0, 1)$, and $\gamma^* = (1, 1, 1, 1, 1)^T$. Following the approach in Zhou and Li (2014), we generate the binary matrix-covariate $\mathbf{C}^* \in 0, 1^{p_1\times p_2}$ as $\mathbf{C}^* = \mathbf{A}_1\mathbf{A}_2^T$, where $\mathbf{A}_1 \in \mathbb{R}^{p_1\times r}$, $\mathbf{A}_2 \in \mathbb{R}^{p_2\times r}$, and r is the rank of $\mathbf{C}^*$. Each entry of $\mathbf{A}_1$ and $\mathbf{A}_2$ which fol-

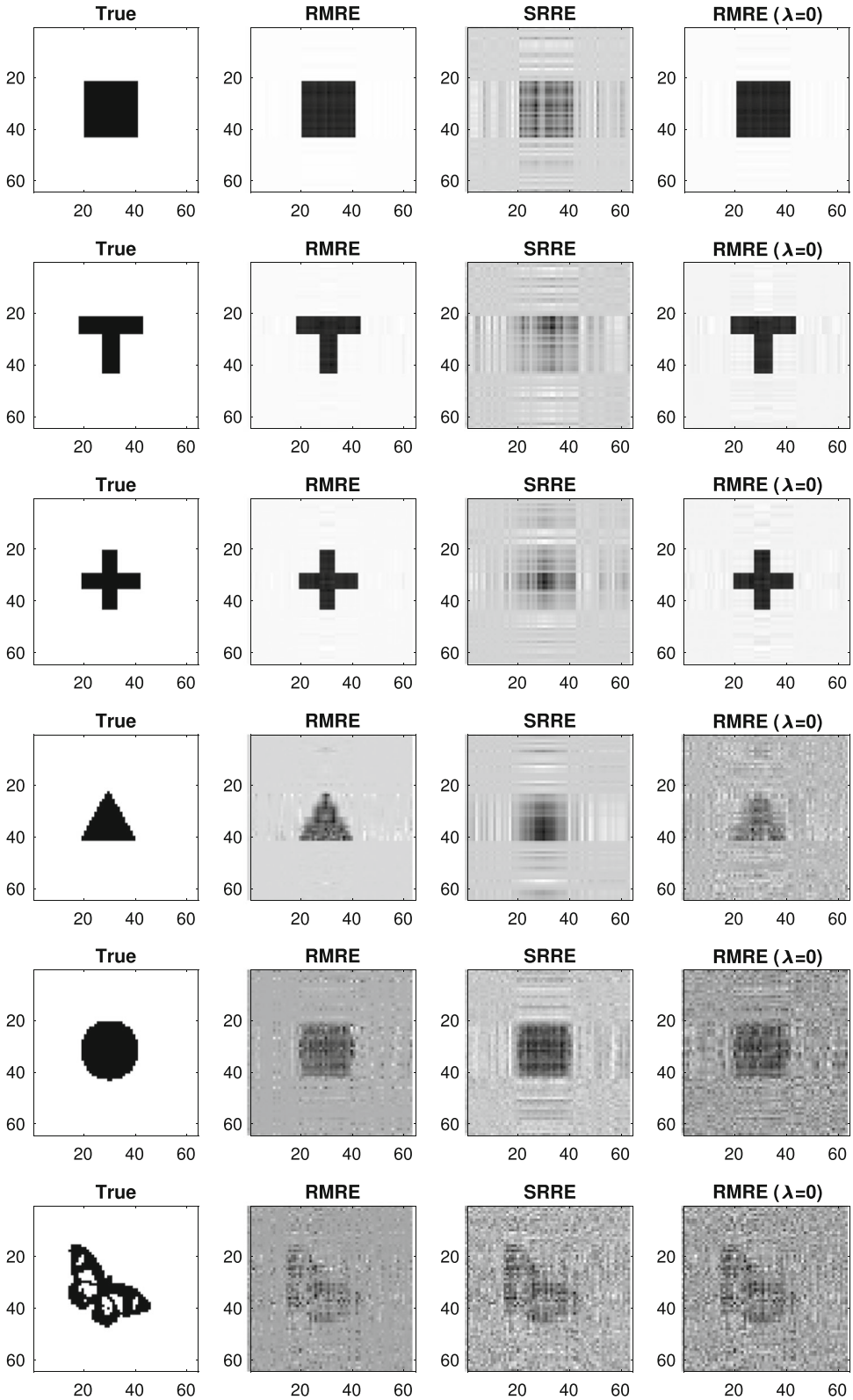

Fig. 1 Comparison of the estimators RMRE and SRRE. The first column are the true signals, the fourth column are recovered signals of RMRE without regularization. The sample size is 500

Table 1 Estimation errors of RMRE, SRRE, and LEME in matrix regression

Shape	Method	$n = 300$	$n = 500$	$n = 700$	$n = 1000$
Square	RMRE	**0.01 (0.00)**	**0.01 (0.00)**	**0.01 (0.00)**	**0.01 (0.00)**
	SRRE	0.19 (0.02)	0.06 (0.01)	0.03 (0.00)	0.02 (0.00)
	LEME	0.06 (0.09)	0.02 (0.03)	0.03 (0.04)	0.03 (0.04)
T	RMRE	**0.05 (0.06)**	**0.01 (0.00)**	**0.01 (0.00)**	**0.01 (0.00)**
	SRRE	0.21 (0.01)	0.14 (0.01)	0.07 (0.01)	0.03 (0.00)
	LEME	0.25 (0.10)	0.05 (0.05)	0.03 (0.00)	0.04 (0.03)
Cross	RMRE	**0.05 (0.06)**	**0.01 (0.01)**	**0.01 (0.00)**	**0.01 (0.00)**
	SRRE	0.19 (0.01)	0.13 (0.01)	0.07 (0.00)	0.03 (0.00)
	LEME	0.23 (0.09)	0.04 (0.03)	0.04 (0.04)	0.02 (0.03)
Triangle	RMRE	**0.18 (0.01)**	**0.10 (0.02)**	**0.06 (0.01)**	**0.06 (0.00)**
	SRRE	**0.18 (0.01)**	0.14 (0.01)	0.12 (0.00)	0.09 (0.00)
	LEME	0.21 (0.04)	0.18 (0.05)	0.12 (0.02)	0.10 (0.01)
Circle	RMRE	0.25 (0.01)	0.15 (0.02)	**0.05 (0.01)**	**0.04 (0.00)**
	SRRE	**0.22 (0.01)**	**0.15 (0.01)**	0.12 (0.01)	0.08 (0.00)
	LEME	0.24 (0.06)	0.18 (0.05)	0.13 (0.00)	0.10 (0.04)
Butterfly	RMRE	0.30 (0.00)	0.27 (0.01)	**0.23 (0.01)**	**0.16 (0.01)**
	SRRE	**0.29 (0.01)**	**0.26 (0.01)**	**0.23 (0.01)**	0.20 (0.01)
	LEME	0.32 (0.02)	0.27 (0.02)	0.26 (0.03)	0.23 (0.03)

The mean (standard deviation) of the RMSEs of $\mathbf{C}^*$ with 100 repetitions are reported. The smallest RMSEs for each setting are highlighted in bold

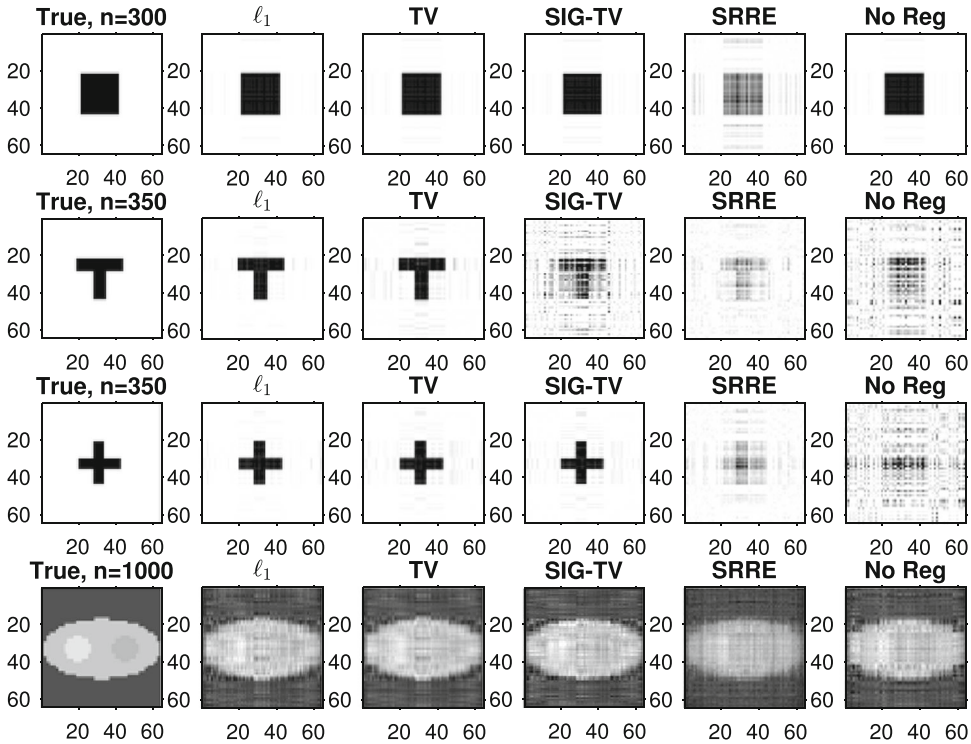

Fig. 2 Comparison of the estimators RMRE, SIG-TV and SRRE. The first column are the true signals with the number of samples is used in each case. The images recovered from RMRE with ℓ_1 regularization and TV regularization are present in the second and the third column respectively

lows a Bernoulli distribution is equal to $\sqrt{1-(1-s)^{1/R}}$ with probability 1, where s controls the sparsity level of $\mathbf{C}^*$. Given the estimation $\hat{\mathbf{C}}$, the prediction $\hat{y}_i$ for $(\mathbf{X}_i, \mathbf{z}_i)$ is defined as $\hat{y} = 1$ if $\log(1 + \exp(\langle \mathbf{X}_i, \mathbf{C} \rangle + \hat{\gamma}^T \mathbf{z}_i)) > 0.5$, and $\hat{y} = 0$ otherwise.

We compare the performance of RMRE with ℓ_1-norm regularization, SRRE, and LEME with various ranks and sparsity levels, and report the prediction errors in Table 2 and the estimation accuracy of $\mathbf{C}^*$ in Table 4. The prediction error is defined as the ratio of mislabeled responses,

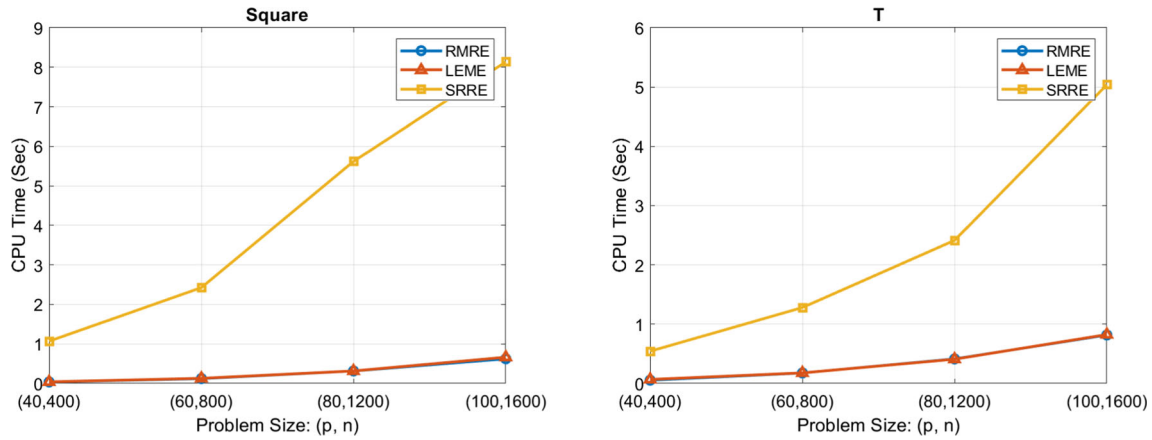

Fig. 3 The computational times of RMRE, SRRE and LEME. The dimension of the matrix covariate is $p \times p$ and the number of samples is n. The computational times of RMRE and LEME are very close

$\sum_{i=1}^{n} |y_i - \hat{y}_i|/n$, and the estimation accuracy is measured by the RMSE of $\hat{\mathbf{C}}$. From Tables 2 and 4, we make the following observations: First, when the signal is sparse or low-rank, our proposed method RMRE demonstrates significantly better performance, which confirms the effectiveness of simultaneously employing regularization and the low-rank constraint in our approach. Second, for small ranks, both RMRE and LEME outperform SRRE, providing evidence that the low-rank constraint enhances the performance of algorithms when the underlying signal is low-rank. Finally, SRRE tends to perform better when the rank is large and the signal is denser, as it does not impose a rank constraint or ℓ_1 regularization.

4.3 Real-world dataset I: EEG

We conducted a comparative analysis of RMRE, SRRE, SDNCMV, and LEME on electroencephalography (EEG) data related to alcoholism (Zhang et al. 1995). The dataset consists of 77 individuals with alcoholism and 45 control individuals (non-alcoholic). During the experiment, subjects were exposed to a stimulus, and voltage values were recorded from 64 channels of electrodes placed on their scalps. The measurements were taken for 256 time points and 120 trials. To derive meaningful insights from this data, we averaged the measurements across the 120 trials, resulting in 122 matrices of size 256×64 for each individual.

The response variable in this study is binary, representing the presence or absence of alcoholism (0 for non-alcoholic and 1 for alcoholic). As the classical linear model is designed for vector-valued covariates, directly vectorizing the matrix data into a high-dimensional vector (e.g., $256 \times 64 = 16,384$ dimensions) may yield poor performance due to the limited sample size of 122. Additionally, vectorization neglects the valuable structural information inherent in the matrix representation of the data (Zhou and Li 2014). To address these challenges, we adopted a fair approach for regularization parameter tuning by performing k-fold cross-validation to divide the entire dataset into training and testing samples. When k equals the sample size, it corresponds to leave-one-out (LOO) cross-validation. Subsequently, within the training data, we applied 5-fold cross-validation to select the optimal shrinkage parameter λ. Finally, we evaluated the performance of the tuned model on the testing data by calculating the misclassification rate.

Through the utilization of this rigorous methodology, our objective is to evaluate the efficacy of RMRE, SRRE, SDNCMV, and LEME in accurately classifying individuals with alcoholism based on EEG data. In Table 5, we report the average misclassification rates along with their corresponding standard deviations across 100 repetitions (reflecting the randomness resulting from the cross-validation procedure's random partitioning). The results in Table 5 indicate that RMRE achieves the smallest misclassification rate among SRRE, SDNCMV, and LEME in the leave-one-out CV, 5-fold CV, and 20-fold CV. Specifically, RMRE achieves a misclassification rate of 0.23, which is slightly higher than the best-performing method, SRRE.

Beyond achieving improved prediction performance, our proposed method provides valuable insights into the underlying structure of the EEG dataset, which SRRE fails to capture. In Fig. 4, we present heatmaps displaying the estimated coefficient matrices obtained using RMRE and SRRE, respectively. The heatmap generated by our method reveals the spatial dependence structure of the predictors, which are sampled from electrodes placed on the scalp, aligning with the electrode locations as defined in the standard electrode position nomenclature (Epstein 2006). In contrast, SRRE and SDNCMV fail to exhibit this spatial-temporal dependence structure. Notably, while LEME incorporates a low-rank structure that induces some spatial dependence among the

Table 2 Prediction errors of RMRE, SRRE, and LEME in the logistic model

Sparsity (%)	Method	Rank			
		$r = 1$	$r = 5$	$r = 10$	$r = 20$
1	RMRE	**0.12 (0.02)**	**0.25 (0.04)**	**0.30 (0.04)**	**0.33 (0.03)**
	SRRE	0.24 (0.03)	0.35 (0.03)	0.36 (0.03)	0.38 (0.03)
	LEME	0.23 (0.08)	0.40 (0.01)	0.40 (0.00)	0.41 (0.01)
5	RMRE	**0.10 (0.02)**	**0.35 (0.03)**	**0.39 (0.03)**	**0.40 (0.02)**
	SRRE	0.20 (0.02)	0.37 (0.03)	0.40 (0.02)	0.41 (0.03)
	LEME	0.18 (0.02)	0.38 (0.01)	0.43 (0.02)	0.43 (0.02)
10	RMRE	**0.12 (0.02)**	**0.37 (0.03)**	0.41 (0.02)	**0.41 (0.02)**
	SRRE	0.20 (0.03)	0.37 (0.03)	0.40 (0.02)	0.41 (0.02)
	LEME	0.13 (0.01)	0.42 (0.00)	**0.40 (0.01)**	0.44 (0.01)
20	RMRE	**0.13 (0.02)**	0.38 (0.03)	0.41 (0.03)	0.41 (0.02)
	SRRE	0.20 (0.03)	**0.35 (0.03)**	**0.38 (0.03)**	**0.40 (0.02)**
	LEME	0.14 (0.04)	0.41 (0.07)	0.44 (0.00)	0.41 (0.02)
50	RMRE	**0.13 (0.02)**	0.36 (0.03)	0.39 (0.02)	0.41 (0.02)
	SRRE	0.19 (0.03)	**0.28 (0.03)**	**0.31 (0.04)**	**0.35 (0.03)**
	LEME	0.16 (0.00)	0.39 (0.04)	0.43 (0.02)	0.45 (0.01)

The mean (standard deviation) of prediction error in $\hat{y}$ with 100 repetitions are reported. The smallest RMSEs for each setting are highlighted in bold

Table 3 Estimation errors of RMRE, SRRE, and LEME in the logistic model

Sparsity (%)	Method	$r = 1$	$r = 5$	$r = 10$	$r = 20$
		Rank			
1	RMRE	0.07 (0.03)	**0.08 (0.02)**	0.09 (0.02)	0.10 (0.01)
	SRRE	0.09 (0.03)	0.09 (0.02)	0.10 (0.02)	0.10 (0.01)
	LEME	**0.07 (0.01)**	0.10 (0.01)	**0.08 (0.01)**	**0.09 (0.00)**
5	RMRE	**0.19 (0.03)**	**0.22 (0.03)**	**0.22 (0.03)**	0.22 (0.02)
	SRRE	0.21 (0.03)	0.23 (0.03)	**0.22 (0.03)**	0.22 (0.02)
	LEME	0.19 (0.09)	0.24 (0.01)	0.23 (0.06)	**0.22 (0.01)**
10	RMRE	**0.28 (0.04)**	0.33 (0.04)	0.33 (0.03)	**0.33 (0.03)**
	SRRE	0.30 (0.05)	0.33 (0.04)	0.33 (0.03)	**0.33 (0.03)**
	LEME	0.29 (0.04)	**0.31 (0.06)**	**0.31 (0.01)**	0.34 (0.04)
20	RMRE	**0.41 (0.04)**	**0.50 (0.04)**	0.52 (0.04)	0.51 (0.04)
	SRRE	0.44 (0.05)	**0.50 (0.04)**	0.52 (0.04)	0.51 (0.04)
	LEME	0.42 (0.05)	0.53 (0.05)	**0.50 (0.02)**	**0.47 (0.03)**
50	RMRE	**0.66 (0.04)**	0.98 (0.07)	**1.03 (0.07)**	1.05 (0.06)
	SRRE	0.69 (0.04)	0.99 (0.07)	**1.03 (0.07)**	1.05 (0.06)
	LEME	0.71 (0.03)	**0.88 (0.06)**	1.09 (0.09)	**1.05 (0.03)**

The mean (standard deviation) of the RMSEs of $\mathbf{C}^*$ with 100 repetitions are reported. The smallest RMSEs for each setting are highlighted in bold

electrodes, it is less apparent compared to our proposed method.

4.4 Real-world dataset II: Leucorrhea

The second experiment involves the classification of IEEE leucorrhea microscopic images (Hao et al. 2019), which are categorized into 6 classes: Erythrocytes (Ery), Leukocytes (Leu), Molds, Epithelial Cells (Epi), and Pyocytes (Pyo). Figure 5 shows some sample images from this dataset. For this experiment, we randomly sample 120 images, with 60 images each from the Leu and Pyo categories. The resolution of each image is downsampled to 32 by 32 pixels.

The performance of the proposed method, RMRE, in classifying the EEG data and the leucorrhea data using leave-one-out, 5-fold, 10-fold, and 20-fold cross-validation is summarized in Table 6. These tables present the mean misclassification rates along with their standard deviations, aver-

Table 4 Estimation errors of RMRE, SRRE, and LEME in the logistic model

Sparsity (%)	Method	$r = 1$	$r = 5$	$r = 10$	$r = 20$
		Rank			
1	RMRE	0.07 (0.03)	**0.08 (0.02)**	0.09 (0.02)	0.10 (0.01)
	SRRE	0.09 (0.03)	0.09 (0.02)	**0.10 (0.02)**	0.10 (0.01)
	LEME	**0.07 (0.01)**	0.10 (0.01)	0.08 (0.01)	**0.09 (0.00)**
5	RMRE	**0.19 (0.03)**	**0.22 (0.03)**	**0.22 (0.03)**	0.22 (0.02)
	SRRE	0.21 (0.03)	0.23 (0.03)	**0.22 (0.03)**	0.22 (0.02)
	LEME	0.19 (0.09)	0.24 (0.01)	0.23 (0.06)	**0.22 (0.01)**
10	RMRE	**0.28 (0.04)**	0.33 (0.04)	0.33 (0.03)	**0.33 (0.03)**
	SRRE	0.30 (0.05)	0.33 (0.04)	0.33 (0.03)	**0.33 (0.03)**
	LEME	0.29 (0.04)	**0.31 (0.06)**	**0.31 (0.01)**	0.34 (0.04)
20	RMRE	**0.41 (0.04)**	**0.50 (0.04)**	0.52 (0.04)	0.51 (0.04)
	SRRE	0.44 (0.05)	**0.50 (0.04)**	0.52 (0.04)	0.51 (0.04)
	LEME	0.42 (0.05)	0.53 (0.05)	**0.50 (0.02)**	**0.47 (0.03)**
50	RMRE	**0.66 (0.04)**	0.98 (0.07)	**1.03 (0.07)**	1.05 (0.06)
	SRRE	0.69 (0.04)	0.99 (0.07)	**1.03 (0.07)**	1.05 (0.06)
	LEME	0.71 (0.03)	**0.88 (0.06)**	1.09 (0.09)	**1.05 (0.03)**

The mean (standard deviation) of the RMSEs of $\mathbf{C}^*$ with 100 repetitions are reported. The smallest RMSEs for each setting are highlighted in bold

Table 5 Misclassification rates (standard deviation in parentheses) of RMRE, SRRE, SDNCMV, and LEME for the EEG dataset

Method	Leave-one-out	5-fold CV	10-fold CV	20-fold CV
RMRE	**0.21**	**0.23 (0.03)**	0.23 (0.02)	**0.22 (0.01)**
SRRE	0.21	0.23 (0.02)	**0.22 (0.02)**	**0.22 (0.01)**
SDNCMV	0.25	0.23 (0.01)	0.23 (0.01)	0.26 (0.01)
LEME	0.25	0.25 (0.02)	0.26 (0.01)	0.24 (0.02)

The smallest misclassification rates for each setting are highlighted in bold

aged over 100 runs (excluding the leave-one-out CV, which has a deterministic nature). The results in Table 6 demonstrate that RMRE achieves the smallest misclassification rates in the 5-fold, 10-fold, and 20-fold CV, outperforming SRRE, SDNCMV, and LEME. However, it does not exhibit good performance in the leave-one-out CV.

4.5 Discussion of experimental results

We have made several general observations from the simulations conducted in Sects. 4.1–4.4:

(1) RMRE demonstrates superior performance over LEME in the 2D shape experiments and performs comparably in the EEG data experiment. We attribute this difference to the regularization choices employed by RMRE and LEME. RMRE utilizes ℓ_1 (Lasso) regularization ($P(\mathbf{C}) = \|\mathbf{C}\|_1$), which promotes sparsity, while LEME uses a smoother regularization of $P(\mathbf{C}) = \|\mathbf{A}\|_F^2 \|\mathbf{B}\|_F^2$ for $\mathbf{C} = \mathbf{A}\mathbf{B}^T$. This proves beneficial in the 2D shape experiments, where the underlying parameters are sparse.

(2) RMRE exhibits faster computational speed compared to SDNCMV, SRRE, and RLRME. This can be explained by the analysis presented in Sect. 2.1, which demonstrates that RMRE incurs a lower computational cost per iteration. SDNCMV is notably slower as it minimizes an objective function similar to (2.4) without the rank constraint, resulting in a larger number of parameters to estimate. In contrast, RMRE and SRRE only estimate a smaller number ($r(q + m - r) + p$) of parameters, making them more efficient, especially when the rank is small.

(3) Both SRRE and RLRME employ nuclear norm-based regularization, which naturally encourages a low-rank structure (Koltchinskii et al. 2011). In comparison, RMRE incorporates both sparse regularization and a low-rank constraint. Our numerical experiments demonstrate that this "double regularization" approach of RMRE outperforms SRRE and RLRME in most scenarios.

(4) The performance of RMRE is found to be robust and not highly sensitive to the choice of the rank r. The estimation errors remain stable over a wide range of rank choices.

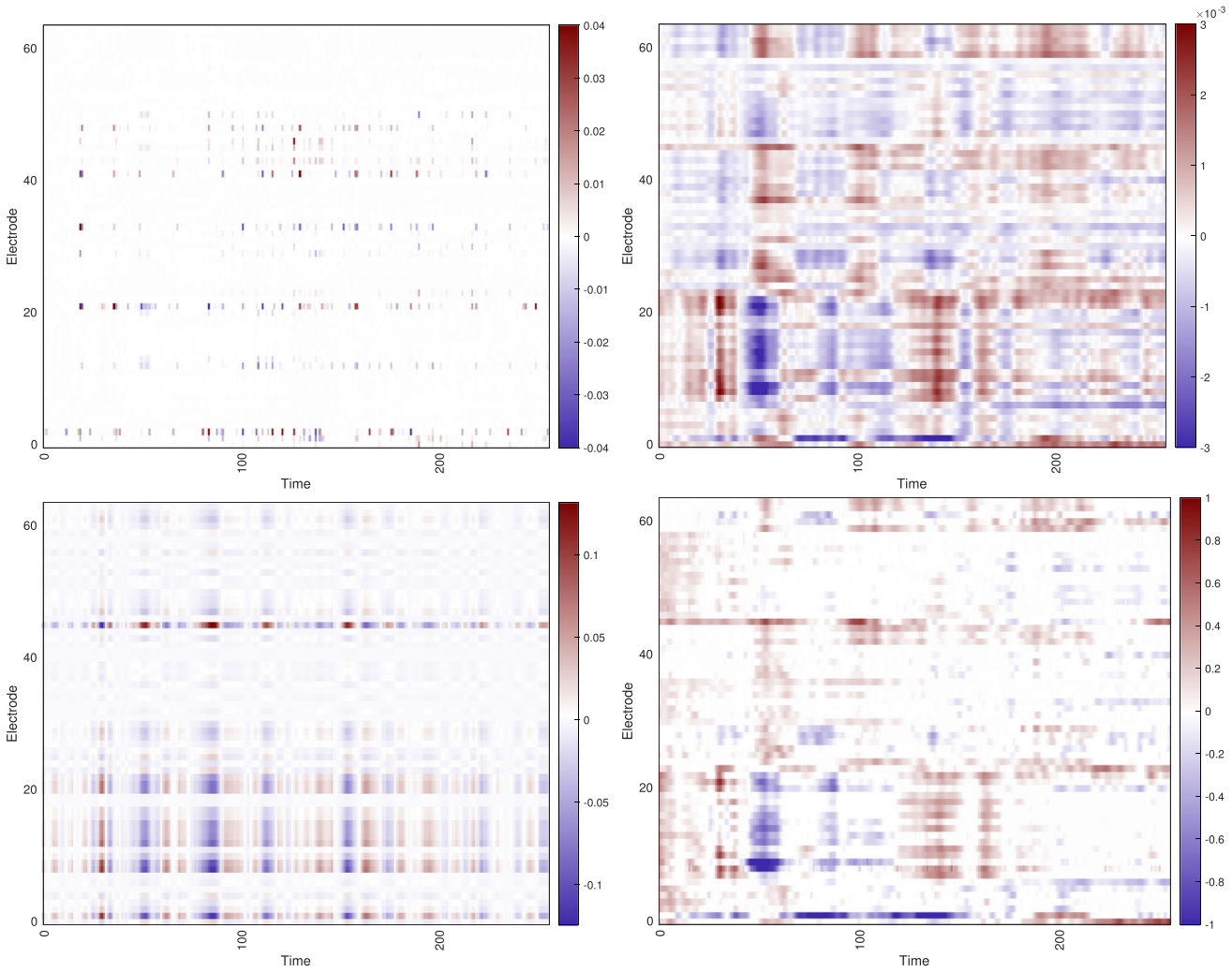

Fig. 4 Heatmaps for $\hat{\mathbf{C}}$ of RMRE (top left), SRRE (top right), LEME (bottom left), and SDNCMV (bottom right) in the EEG dataset

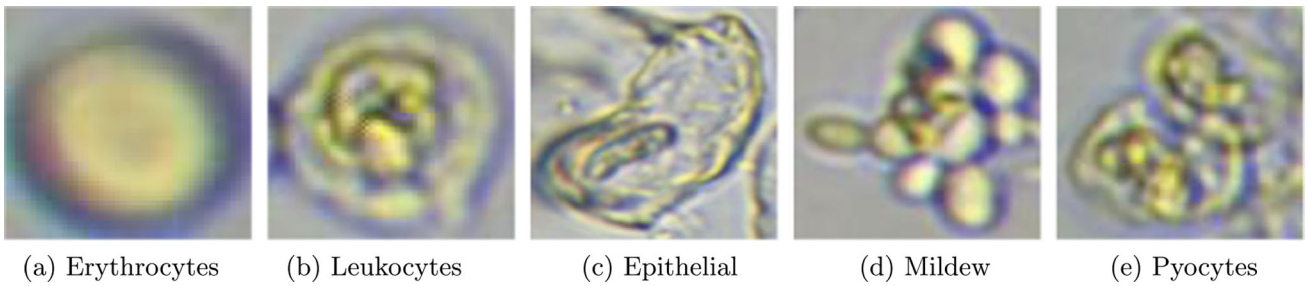

Fig. 5 Microscopic images for each category

Table 6 Misclassification rates (standard deviation in parentheses) of RMRE, SRRE, SDNCMV, and LEME for the leucorrhea data

Method	Leave-one-out	5-fold CV	10-fold CV	20-fold CV
RMRE	0.28	**0.23 (0.02)**	**0.25 (0.02)**	**0.25 (0.02)**
SRRE	0.26	0.26 (0.03)	0.26 (0.01)	0.26 (0.02)
SDNCMV	0.26	0.26 (0.01)	**0.25 (0.01)**	0.27 (0.01)
LEME	**0.25**	0.26 (0.02)	0.26 (0.03)	**0.25 (0.02)**

The smallest misclassification rates for each setting are highlighted in bold

5 Conclusion

In this paper, we have proposed a comprehensive framework for matrix-covariate regression models, offering a versatile approach to handle a variety of response variables. Our method incorporates a general regularization function, allowing for the application of specific techniques such as the lasso, total variation (TV), and fused lasso penalties in practical scenarios. By leveraging a regularization-based objective function and a low-rank constrained optimization approach, our framework stands out from existing methods.

Moreover, we have developed an efficient Riemannian-steepest-descent algorithm and provided rigorous theoretical analysis. Our algorithm guarantees convergence, and we have shown that all accumulation points of the iterates exhibit estimation errors in the order of $O(1/\sqrt{n})$, effectively attaining the minimax rate. Extensive numerical studies have substantiated the advantages of our algorithm, particularly in cases where the underlying signals exhibit both low-rank and sparse structures. These promising results highlight the efficacy and applicability of our proposed framework in matrix-covariate regression problems. Future research directions may involve exploring specific regularization techniques and extending the framework to accommodate other types of response variables.

Acknowledgements The authors would like to thank the Editor, the Associate Editor and the reviewers for their constructive and insightful comments that greatly improved the manuscript. This work was partially supported by NSF grants (DMS-1924792, DMS-2318925 and CNS-1818500).

Author Contributions H-HH: conceptualization, methodology, formal analysis, investigation, writing-original draft preparation, writing-review, supervision. FY: conceptualization, methodology, formal analysis, investigation, writing review, programming and numerical results. XF: programming and numerical results. TZ: conceptualization, methodology, formal analysis, investigation, writing-review, supervision. All authors have read and agreed to the published version of the manuscript.

Declarations

Conflict of interest The authors declare no conflict of interest.

Appendix A: Technique proofs

A.1 Proof of Proposition 1

Proof (a) By the line search rule, we have that $F(\mathbf{C}^{(k+1)}, \gamma^{(k+1)}) \leq F(\mathbf{C}^{(\text{iter})}, \gamma^{(\text{iter})})$ for all $k \geq 1$. Since F is bounded below, the limit $\lim_{k\to\infty} F(\mathbf{C}^{(\text{iter})}, \gamma^{(\text{iter})})$ exists. Assume that one of the limiting point of the sequence $(\mathbf{C}^{(\text{iter})}, \gamma^{(\text{iter})})$ is $(\tilde{\mathbf{C}}, \tilde{\gamma})$, then the line search rule implies that $\frac{\partial}{\partial_\gamma} F(\tilde{\mathbf{C}}, \tilde{\gamma}) = 0$ and $\frac{\partial}{\partial \mathbf{C}} F(\tilde{\mathbf{C}}, \tilde{\gamma}) = 0$.

(b) The proof follows (Absil et al. 2009, Theorem 4.3.1). □

A.2 Proof of Lemma 5

Proof We assume that for any $\mathbf{x}$ in the neighborhood of $\mathbf{x}^*$, $\mathbf{x} - \mathbf{x}^*$ can be uniquely decomposed into $\mathbf{x} - \mathbf{x}^* = \mathbf{x}^{(1)} + \mathbf{x}^{(2)}$ such that $\mathbf{x}^{(1)} \in T_{\mathbf{x}^*}(\mathcal{M})$ and $\mathbf{x}^{(2)} \in T_{\mathbf{x}^*,\perp}(\mathcal{M})$. Let $b = \|\mathbf{x} - \mathbf{x}^*\|$, then if $b \leq c_0$, then $\|\mathbf{x}^{(1)}\| \leq b$ and $\|\mathbf{x}^{(2)}\| \leq C_T b^2$.

Let $\mathbf{v} = \frac{\mathbf{x}-\mathbf{x}^*}{\|\mathbf{x}-\mathbf{x}^*\|}$ be the direction from $\mathbf{x}^*$ to $\mathbf{x}$, then

$$f(\mathbf{x}) - f(\mathbf{x}^*) = \int_{\mathbf{x}^*}^{\mathbf{x}} \langle \mathbf{v}, \nabla f(\mathbf{t}) \rangle \mathrm{d}\mathbf{t} = \langle \mathbf{x} - \mathbf{x}^*, \nabla f(\mathbf{x}^*) \rangle + \int_{\mathbf{x}^*}^{\mathbf{x}} \langle \mathbf{v}, \nabla f(\mathbf{t}) - \nabla f(\mathbf{x}^*) \rangle \mathrm{d}\mathbf{t},$$

where the first term can be bounded by

$$\begin{aligned}\langle \mathbf{x} - \mathbf{x}^*, \nabla f(\mathbf{x}^*) \rangle &= \langle \mathbf{x}^{(1)} + \mathbf{x}^{(2)}, \nabla f(\mathbf{x}^*) \rangle \\ &= \langle \mathbf{x}^{(1)}, \Pi_{T_{\mathbf{x}^*}(\mathcal{M})} \nabla f(\mathbf{x}^*) \rangle + \langle \mathbf{x}^{(2)}, \Pi_{T_{\mathbf{x}^*}(\mathcal{M}),\perp} \nabla f(\mathbf{x}^*) \rangle \\ &\leq b \|\Pi_{T_{\mathbf{x}^*}(\mathcal{M})} \nabla f(\mathbf{x}^*)\| + C_T b^2 \|\Pi_{T_{\mathbf{x}^*}(\mathcal{M}),\perp} \nabla f(\mathbf{x}^*)\|.\end{aligned}$$

On the other hand, the second term can be bounded by

$$\int_{\mathbf{x}^*}^{\mathbf{x}} \langle \mathbf{v}, \nabla f(\mathbf{t}) - \nabla f(\mathbf{x}^*) \rangle \mathrm{d}\mathbf{t} \geq \frac{1}{2} b^2 C_{H,1}.$$

Combining these two inequalities, the lemma is proved. □

A.3 Technical Assumptions

We first present a few conditions for establishing the model consistency of the proposed estimator in (2.4).

Assumption 1 There exists a positive constant $C_1 > 0$ such that

$$\frac{1}{n} \Big\| \sum_{i=1}^{n} \mathrm{vec}(\mathbf{X}_i, \mathbf{z}_i) \mathrm{vec}(\mathbf{X}_i, \mathbf{z}_i)^T \Big\| \leq C_1,$$

where $\mathrm{vec}(\mathbf{X}_i, \mathbf{z}_i) \in \mathbb{R}^{qm+p}$ is a vector consisting of the elements in $\mathbf{X}_i \in \mathbb{R}^{m\times q}$ and $\mathbf{z}_i \in \mathbb{R}^p$.

Assumption 2 For $\mathbf{H}(\mathbf{C}, \gamma) \in \mathbb{R}^{(qm+p)\times(qm+p)}$ defined by

$$\mathbf{H}(\mathbf{C}, \gamma) = \sum_{i=1}^{n} w_{2,i} \mathrm{vec}(\mathbf{X}_i, \mathbf{z}_i) \mathrm{vec}(\mathbf{X}_i, \mathbf{z}_i)^T, \tag{A1}$$

where

$$w_{2,i} = \begin{cases} 2, & \text{for the ordinary matrix} \\ & \text{-covariate regression model,} \\ \frac{e^{\langle \mathbf{X}_i, \mathbf{C} \rangle + \mathbf{z}_i^T \gamma}}{(1+e^{\langle \mathbf{X}_i, \mathbf{C} \rangle + \mathbf{z}_i^T \gamma})^2}, & \text{for the logistic matrix} \\ & \text{-covariate regression model,} \end{cases}$$

and $\frac{1}{n}\mathbf{H}(\mathbf{C},\gamma)$ is positive definite with eigenvalues bounded from below for all $(\mathbf{C},\gamma)$ in a neighborhood of $(\mathbf{C}^*,\gamma^*)$. Specifically, there exists positive constants $C_2 > 0$ and $c_0 \leq \sigma_r(\mathbf{C}^*)/2$ such that $\frac{1}{n}\lambda_{\min}\Big(\mathbf{H}(\mathbf{C},\gamma)\Big) \geq C_2$ and for all $(\mathbf{C},\gamma)$ such that

$$\text{dist}((\mathbf{C},\gamma),(\mathbf{C}^*,\gamma^*)) = \sqrt{\|\mathbf{C}-\mathbf{C}^*\|_F^2 + \|\gamma-\gamma^*\|^2} \leq c_0.$$

Assumptions 1 and 2 can be considered as the generalized versions of the restricted isometry property (RIP) (Candes and Tao 2005) in Recht et al. (2010) and are comparable to (Li et al. 2021, Conditions 1,2,5,6). In particular, Assumption 1 ensures that the sensing vectors $\text{vec}(\mathbf{X}_i,\mathbf{z}_i) \in \mathbb{R}^{qm+p}$ are not the same or in a similar direction, and it is satisfied if the sensing vectors are sampled from a distribution that is relatively uniformly distributed among all directions. Assumption 2 ensures that the Hessian matrix of $\sum_{i=1}^n l(y_i,\langle\mathbf{X}_i,\mathbf{C}\rangle+\gamma^T\mathbf{z}_i),\mathbf{H}$, is not a singular matrix.

Remark 2 Assumptions 1 and 2 are reasonable when $n > O(pm+q)$ and $(\mathbf{X}_i,\mathbf{z}_i)$ are sampled from a reasonable distribution that does not concentrate around certain directions. For example, if $\text{vec}(\mathbf{X}_i,\mathbf{z}_i)$ are i.i.d. sampled from a distribution of $N(0,\mathbf{I})$, then the standard concentration of measure results (Wainwright 2019) imply that for the matrix-covariate regression model, $\mathbf{H} = 2\sum_{i=1}^n \text{vec}(\mathbf{X}_i,\mathbf{z}_i)\text{vec}(\mathbf{X}_i,\mathbf{z}_i)^T$, and the standard concentration of measure results (Wainwright 2019) imply that Assumption 1 holds with a high probability when $n = O(qm+p)$ and Assumption 2 holds for the regression model as well since $\mathbf{H} = 2\sum_{i=1}^n \text{vec}(\mathbf{X}_i,\mathbf{z}_i)\text{vec}(\mathbf{X}_i,\mathbf{z}_i)^T$. Assumption 2 holds for the logistic regression model as long as $a_i = \langle\mathbf{X}_i,\mathbf{C}\rangle + \mathbf{z}_i^T\gamma$ is bounded above for most indices i as $w_{2,i} \geq 1/(2e^{a_i})$.

Assumption 3 (Assumption on the noise for the matrix-covariate regression model): For the matrix-covariate regression model, the error ϵ_i's in Eq. (2.1) follow an independent and identically distributed (i.i.d.) zero-mean and sub-Gaussian distributions with zero mean and variance one, i.e., $\text{E}(\epsilon_i) = 0$ and $\text{Var}(\epsilon_i) = 1$ to ensure that the distribution of outliers is limited as this model is not designed to handle outliers.

Assumption 4 There exists a positive constant $C_3 > 0$ such that $\frac{1}{n}\Big\|\mathbf{H}(\mathbf{C},\gamma)\Big\| \leq C_3$ for all $(\mathbf{C},\gamma)$ such that $\text{dist}((\mathbf{C},\gamma),(\mathbf{C}^*,\gamma^*)) = \sqrt{\|\mathbf{C}-\mathbf{C}^*\|_F^2 + \|\gamma-\gamma^*\|^2} \leq c_0$. Combining it with Assumption 2, it suggests that $\mathbf{H}$ is a well-conditioned matrix. Hence, Assumption 4 can be considered as the generalized version of the Restricted Isometry condition in Recht et al. (2010) and comparable to Conditions 2,6 of (Li et al. 2021).

Assumptions 4 is reasonable when $n > O(pm+q)$ and $(\mathbf{X}_i,\mathbf{z}_i)$ is sampled from a reasonable distribution that does not concentrate around certain directions. For example, if $\text{vec}(\mathbf{X}_i,\mathbf{z}_i)$ are i.i.d. sampled from a distribution of $N(0,\mathbf{I})$, then the standard concentration of measure results (Wainwright 2019) imply that Assumption 2 holds for all three models with a high probability, since $w_{2,i}$ are bounded above. With Assumption 4, we have the following result showing the convergence of Algorithm 1, with its proof deferred to the appendix. It shows that with a good initialization, all accumulation points have estimation errors converging to zero as $n \to \infty$.

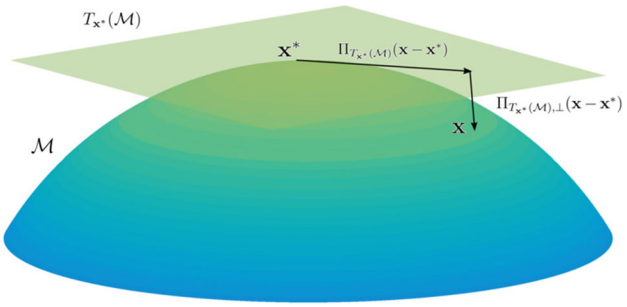

Fig. 6 A visualization of the manifold $\mathcal{M}$, two points $\mathbf{x},\mathbf{x}^* \in \mathcal{M}$, the tangent space $T_{\mathbf{x}^*}(\mathcal{M})$, and the projectors $\Pi_{T_{\mathbf{x}^*}(\mathcal{M})}$ and $\Pi_{T_{\mathbf{x}^*}(\mathcal{M}),\perp}$

We will need the following assumptions:

Assumption 5 There exists $C_{upper} > 0$ such that for all $(\mathbf{C},\gamma) \in \mathcal{A}(r,a)$,

$$\sum_{i=1}^n \Big(\langle\mathbf{X}_i,\mathbf{C}\rangle + \gamma^T\mathbf{z}_i\Big)^2 \leq C_{upper} n\|\text{vec}(\mathbf{C},\gamma)\|^2.$$

Assumption 6 There exists $c_\epsilon > 0$ such that for all $x \in \mathbb{R}$, $KL(P_{\epsilon,0},P_{\epsilon,x}) \leq c_\epsilon x^2$, where $P_{\epsilon,x}$ represent the distribution of $\epsilon_i + x$ for the matrix-covariate regression with continuous responses models, and $P_{\epsilon,x}$ represent the Bernoulli distribution with parameter x for the logistic regression model with binary responses. In addition, KL represents the Kullback–Leibler divergence: For distributions P and Q of a continuous random variable with probability density functions $p(x)$ and $q(x)$, it is defined to be the integral $KL(P,Q) = \int_x p(x)\log(p(x)/q(x))\text{d}x$.

Assumption 5 is less restrictive of Assumption 1 as it only needs to be true for all $(\mathbf{C},\gamma) \in \mathcal{A}(r,a)$, and Assumption 6 holds under both the matrix-covariate logistic regression model for binary responses and matrix-covariate regression models for continuous responses, Assumption 6 holds for zero-mean, symmetric distributions with tails decaying not faster than Gaussian, including Gaussian distribution, exponential distribution, Cauchy distribution, Bernoulli distribution, and Student's t distribution.

A.4 Sketch of the Proof of Theorem 2

We start with the intuition of the proof with a function $f : \mathbb{R}^p \to \mathbb{R}$. To show that f has a local minimizer around $\mathbf{x}^*$, it is sufficient to show that the gradient $\nabla f(\mathbf{x}^*) \approx 0$ and the Hessian matrix of $f(\mathbf{x})$, $\mathbf{H}(\mathbf{x})$, is positive definite with eigenvalues strictly larger than some constant $c > 0$. The intuition of the proof follows from the Taylor expansion that

$$\begin{aligned} f(\mathbf{x}) \approx & f(\mathbf{x}^*) + (\mathbf{x}-\mathbf{x}^*)^T \nabla f(\mathbf{x}^*) + \frac{1}{2}(\mathbf{x}-\mathbf{x}^*)^T H(\mathbf{x}^*)(\mathbf{x}-\mathbf{x}^*) \\ \geq & f(\mathbf{x}^*) + (\mathbf{x}-\mathbf{x}^*)^T \nabla f(\mathbf{x}^*) + \frac{c}{2}\|\mathbf{x}-\mathbf{x}^*\|^2. \end{aligned} \tag{A2}$$

As a result, there is local minimizer in the neighbor of $\mathbf{x}^*$ with radius $\frac{\|\nabla f(\mathbf{x}^*)\|}{c}$, i.e., $B\left(\mathbf{x}^*, \frac{\|\nabla f(\mathbf{x}^*)\|}{c}\right)$. To extend this proof to (2.4), the main obstacle is the nonlinear constraint in the optimization problem. To address this issue, we consider the constraint set in (2.4) as a manifold and generalize the "second-order Taylor expansion" in (A2) to the function defined on a manifold. With this generalized Taylor expansion, a similar strategy can be applied to prove that the minimizer of (2.4) is close to $(\mathbf{C}^*, \gamma^*)$.

To analyze functions defined on manifolds, we introduce a few additional notations. We assume a manifold $\mathcal{M} \subset \mathbb{R}^p$ and a function $f : \mathbb{R}^p \to \mathbb{R}$, and investigate $f(\mathbf{x})$ for $\mathbf{x} \in B(\mathbf{x}^*, r) \cap \mathcal{M}$, i.e., a local neighborhood of $\mathbf{x}^*$ on the manifold $\mathcal{M}$. We denote the first and second derivatives of $f(\mathbf{x})$ by $\nabla f(\mathbf{x}) \in \mathbb{R}^p$ and $\mathbf{H}(\mathbf{x}) \in \mathbb{R}^{p\times p}$ respectively, the tangent plane of $\mathcal{M}$ at $\mathbf{x}^*$ by $T_{\mathbf{x}^*}(\mathcal{M})$, and let $\Pi_{T_{\mathbf{x}^*}(\mathcal{M})}$ and $\Pi_{T_{\mathbf{x}^*}(\mathcal{M}),\perp}$ be the projectors to $T_{\mathbf{x}^*}(\mathcal{M})$ and its orthogonal subspace respectively. These definitions are visualized in Fig. 6.

Then, we say that a manifold $\mathcal{M}$ is curved with parameter (c_0, C_T) at $\mathbf{x}^* \in \mathcal{M}$, if for any $\mathbf{x} \in B(\mathbf{x}^*, c_0) \cap \mathcal{M}$, we have $\|\Pi_{T_{\mathbf{x}^*}(\mathcal{M}),\perp}(\mathbf{x}-\mathbf{x}^*)\| \leq C_T\|\Pi_{T_{\mathbf{x}^*}(\mathcal{M})}(\mathbf{x}-\mathbf{x}^*)\|^2$. Intuitively, it means that the projection of $\mathbf{x}-\mathbf{x}^*$ to the tangent space $T_{\mathbf{x}^*}(\mathcal{M})$ has a larger magnitude than the projection to the orthogonal subspace of the tangent space (see Fig. 6).We remark that a larger C_T means that the manifold $\mathcal{M}$ is more "curved" around $\mathbf{x}^*$. Then, Lemma 5 establishes the lower bound of f based on the local properties such as the first and second derivatives of f at $\mathbf{x}^*$, the tangent space of $\mathcal{M}$ around $\mathbf{x}^*$, and the curvature parameters (c_0, C_T).

Lemma 5 *Consider a d-dimensional manifold $\mathcal{M} \subset \mathbb{R}^p$ and a function $f : \mathbb{R}^P \to \mathbb{R}$, define $C_{H,1} = \min_{\mathbf{x}\in B(\mathbf{x}^*,c_0)} \lambda_{\min}(\mathbf{H}(\mathbf{x}))$, and assume that $\mathcal{M}$ is curved with parameter (c_0, C_T) at $\mathbf{x}^*$ and $4C_{H,1} \geq C_T\|\Pi_{T_{\mathbf{x}^*}(\mathcal{M}),\perp}\nabla f(\mathbf{x}^*)\|$, then we have the following lower bound for any $\mathbf{x} \in B(\mathbf{x}^*, c_0) \cap \mathcal{M}$:*

$$\begin{aligned} f(\mathbf{x}) - f(\mathbf{x}^*) \geq & \frac{1}{2}b^2 C_{H,1} - b\|\Pi_{T_{\mathbf{x}^*}(\mathcal{M})}\nabla f(\mathbf{x}^*)\| \\ & - C_T b^2 \|\Pi_{T_{\mathbf{x}^*}(\mathcal{M}),\perp}\nabla f(\mathbf{x}^*)\|, \end{aligned}$$

where $b = \|\mathbf{x}-\mathbf{x}^*\|$.

Lemma 5 can be viewed as a generalization of Inequality (A2): when $\mathcal{M} = \mathbb{R}^p$, then we have $C_T = 0$, $\Pi_{T,\perp} = \emptyset$ and as a result, $\|\Pi_{T_{\mathbf{x}^*}(\mathcal{M}),\perp}\nabla f(\mathbf{x}^*)\| = 0$. To apply Lemma 5 to our problem, we need to estimate the parameters $c_0, C_T, C_{H,1}$, $\|\Pi_{T_{\mathbf{x}^*}(\mathcal{M})}\nabla f(\mathbf{x}^*)\|$, and $\|\Pi_{T_{\mathbf{x}^*}(\mathcal{M}),\perp}\nabla f(\mathbf{x}^*)\|$ in the statement of Lemma 5. In particular, (c_0, C_T) depends on the manifold $\mathcal{M}$ used in the optimization problem (2.4), which is

$$\mathcal{M} = \{(\mathbf{C}, \gamma) \in \mathbb{R}^{q\times m} \times \mathbb{R}^p : \mathrm{rank}(\mathbf{C}) = r\}. \tag{A3}$$

By treating $\mathcal{M}$ as the product space of $\mathbb{R}^p$ and the manifold of low-rank matrices $\{\mathbf{C} \in \mathbb{R}^{q\times m} : \mathrm{rank}(\mathbf{C}) = r\}$ and following the tangent space of the set of low-rank matrices in the literature (Absil and Oseledets 2015; Zhang and Yang 2018), we obtain the following lemma of "curvedness" parameters (c_0, C_T) of $\mathcal{M}$ at $(\mathbf{C}^*, \gamma^*)$.

Lemma 6 *The manifold $\mathcal{M}$ defined in (A3) is curved with parameter (c_0, C_T) at $(\mathbf{C}^*, \gamma^*)$, for any $c_0 \leq \sigma_r(\mathbf{C}^*)/2$ and $C_T = 2/\sigma_r(\mathbf{C}^*)$, where $\sigma_r(\mathbf{C}^*)$ represents the r-th (i.e., the smallest) singular value of $\mathbf{C}^*$.*

In addition, the parameter $C_{H,1}$ in Lemma 5 can be estimated from Assumption 2, and the derivatives $\|\Pi_{T_{\mathbf{x}^*}(\mathcal{M})}\nabla f(\mathbf{x}^*)\|$ and $\|\Pi_{T_{\mathbf{x}^*}(\mathcal{M}),\perp}\nabla f(\mathbf{x}^*)\|$ in Lemma 5 can be estimated from Assumptions 1 and 3. Then the proof of Theorem 2 follows from Lemma 5 and the intuition introduced at the beginning of this section, with technical details deferred to the appendix.

A.5 Proof of Lemma 6

Proof By following the tangent space of the set of low-rank matrices in the literature (Absil and Oseledets 2015; Zhang and Yang 2018), we have the following explicit expressions of the tangent space of $\mathcal{M}$ at $(\mathbf{C}^*, \gamma^*)$ that

$$T_{(\mathbf{C}^*,\gamma^*),\mathcal{M}} = \{(\mathbf{A}\mathbf{V}^{*T} + \mathbf{U}^*\mathbf{U}^{*T}\mathbf{B}, \mathbf{y}) : \mathbf{A}, \mathbf{B} \in \mathbb{R}^{q\times m}, \mathbf{y} \in \mathbb{R}^p\},$$

where $\mathbf{U}^* \in \mathbb{R}^{q\times r}$ and $\mathbf{V}^* \in \mathbb{R}^{m\times r}$ are obtained from the singular value decomposition of $\mathbf{C}^*$ such that $\mathbf{C}^* = \mathbf{U}^*\mathbf{\Sigma}\mathbf{V}^{*T}$. The projection operators in Lemma 5 are given by

$$\begin{aligned} &\Pi_{T_{(\mathbf{C}^*,\gamma^*),\mathcal{M}}}(\mathbf{D}, \mathbf{y}) = (\mathbf{U}^*\mathbf{U}^{*T}\mathbf{D} + \mathbf{D}\mathbf{V}^*\mathbf{V}^{*T} - \mathbf{U}^*\mathbf{U}^{*T}\mathbf{D}\mathbf{V}^*\mathbf{V}^{*T}, \mathbf{y}), \\ &\Pi_{T_{(\mathbf{C}^*,\gamma^*),\mathcal{M}},\perp}(\mathbf{D}, \mathbf{y}) = ((\mathbf{I} - \mathbf{U}^*\mathbf{U}^{*T})\mathbf{D}(\mathbf{I} - \mathbf{V}^*\mathbf{V}^{*T}), 0). \end{aligned}$$

By choosing $\mathbf{U}^{*\perp} \in \mathbb{R}^{q\times(q-r)}$ such that $[\mathbf{U}^{*\perp}, \mathbf{U}^*] \in \mathbb{R}^{q\times q}$ is orthogonal and choose $\mathbf{V}^{*\perp} \in \mathbb{R}^{m\times(m-r)}$ such that $[\mathbf{V}^{*\perp}, \mathbf{V}^*] \in \mathbb{R}^{m\times m}$ is orthogonal, then we can express the projectors as follows: for any $(\mathbf{C}, \gamma)$ close to $(\mathbf{C}^*, \gamma^*)$, we may write $\mathbf{C}-\mathbf{C}^* = \mathbf{U}^*\mathbf{D}_1\mathbf{V}^{*T} + \mathbf{U}^{*\perp}\mathbf{D}_2\mathbf{V}^{*T} + \mathbf{U}^*\mathbf{D}_3\mathbf{V}^{*\perp T} +$

$$\mathbf{U}^{*\perp}\mathbf{D}_4\mathbf{V}^{*\perp T},$$

$$\|\Pi_{T_{(\mathbf{C}^*,\gamma^*),\mathcal{M},\perp}}(\mathbf{C}-\mathbf{C}^*,\gamma-\gamma^*)\| = \|(\mathbf{U}^{*\perp T}\mathbf{D}_4\mathbf{V}^{*\perp},0)\| = \|\mathbf{D}_4\|_F,$$

and

$$\|\Pi_{T_{(\mathbf{C}^*,\gamma^*),\mathcal{M}}}(\mathbf{C}-\mathbf{C}^*,\gamma-\gamma^*)\| = \sqrt{\|\mathbf{D}_1\|_F^2+\|\mathbf{D}_2\|_F^2+\|\mathbf{D}_3\|_F^2+\|\gamma-\gamma^*\|^2}.$$

By $\text{rank}(\mathbf{D}) = r$, we have $\mathbf{D}_4 = \mathbf{D}_2(\mathbf{D}_1+\mathbf{U}^{*T}\mathbf{C}^*\mathbf{V}^*)^{-1}\mathbf{D}_3$. Thus, when $\|\mathbf{C}-\mathbf{C}^*\|_F \le \sigma_r(\mathbf{C}^*)/2$,

$$\begin{aligned}&\|\Pi_{T_{(\mathbf{C}^*,\gamma^*),\mathcal{M},\perp}}(\mathbf{C}-\mathbf{C}^*,\gamma-\gamma^*)\|\\ &\le \frac{\|\mathbf{D}_2\|_F\|\mathbf{D}_3\|_F}{\sigma_r(\mathbf{D}_1)-\|\mathbf{C}-\mathbf{C}^*\|_F}\\ &\le \frac{2\|\Pi_{T_{(\mathbf{C}^*,\gamma^*),\mathcal{M}}}(\mathbf{C}-\mathbf{C}^*,\gamma-\gamma^*)\|^2}{\sigma_r(\mathbf{C}^*)},\end{aligned}$$

and Lemma 6 is proved. □

A.6 Proof of Theorem 2

Proof In the proof, we mainly work with

$$f(\mathbf{C},\gamma) = \sum_{i=1}^n l(y_i,\langle\mathbf{X}_i,\mathbf{C}\rangle+\gamma^T\mathbf{z}_i),$$

and it is sufficient to show that for all $(\mathbf{C},\gamma)\in\mathcal{M}$ such that

$$\begin{aligned}\sqrt{\|\mathbf{C}-\mathbf{C}^*\|_F^2+\|\gamma-\gamma^*\|^2} &\ge C_{error,1},\\ f(\mathbf{C},\gamma)-f(\mathbf{C}^*,\gamma^*) &\ge \lambda P(\mathbf{C}^*,\gamma^*).\end{aligned}$$

To prove it, we first calculate the constants and the operators in Lemma 5 as follows. For all three models, the constant on the curvature of $\mathcal{M}$ is the same. Hence, we may choose $C_T = 2/\sigma_{\min}(\mathbf{C}^*)$. In addition, as discussed in the proof of Lemma 6, the projectors Π_T and $\Pi_{T,\perp}$ at $(\mathbf{C}^*,\gamma^*)$ can be defined by

$$\begin{aligned}\Pi_{T_{(\mathbf{C}^*,\gamma^*)}(\mathcal{M})}(\mathbf{C},\gamma) &= (\mathbf{C}-\Pi_{\mathbf{U}^*,\perp}\mathbf{C}\Pi_{\mathbf{V}^*,\perp},\gamma),\\ \Pi_{T_{(\mathbf{C}^*,\gamma^*)}(\mathcal{M}),\perp}(\mathbf{C},\gamma) &= (\Pi_{\mathbf{U}^*,\perp}\mathbf{C}\Pi_{\mathbf{V}^*,\perp},0),\end{aligned} \tag{A4}$$

where $\mathbf{U}^*\in\mathbb{R}^{q\times r}$ and $\mathbf{V}^*\in\mathbb{R}^{m\times r}$ are the left and right singular components of $\mathbf{C}^*$, $\Pi_{\mathbf{U}^*} = \mathbf{U}^*\mathbf{U}^{*T}$, $\Pi_{\mathbf{U}^*,\perp} = \mathbf{I}-\Pi_{\mathbf{U}^*}$, $\Pi_{\mathbf{V}^*} = \mathbf{V}^*\mathbf{V}^{*T}$, and $\Pi_{\mathbf{U}^*,\perp} = \mathbf{I}-\mathbf{V}^*\mathbf{V}^{*T}$. As for the first derivative, we have

$$\nabla f(\mathbf{C}^*,\gamma^*) = \sum_{i=1}^n w_{1,i}\text{vec}(\mathbf{X}_i,\mathbf{z}_i),$$

where

$$w_{1,i} = \begin{cases}2\epsilon_i, & \text{for the matrix variate regression model;}\\ \epsilon_i, & \text{for the logistic matrix variate regression model.}\end{cases}$$

Combining it with (A4),

$$\begin{aligned}\Pi_{T_{(\mathbf{C}^*,\gamma^*)}(\mathcal{M})}\nabla f(\mathbf{C}^*,\gamma^*) &= (\Pi_{\mathbf{U}}\mathbf{X}_i+\mathbf{X}_i\Pi_{\mathbf{V}}-\Pi_{\mathbf{U}}\mathbf{X}_i\Pi_{\mathbf{V}},\mathbf{z}_i),\\ \Pi_{T_{(\mathbf{C}^*,\gamma^*)}(\mathcal{M}),\perp}\nabla f(\mathbf{C}^*,\gamma^*) &= (\mathbf{X}_i-\Pi_{\mathbf{U}}\mathbf{X}_i-\mathbf{X}_i\Pi_{\mathbf{V}}+\Pi_{\mathbf{U}}\mathbf{X}_i\Pi_{\mathbf{V}},0).\end{aligned}$$

Now let us introduce a lemma as follows.

Lemma 7 *For any projection matrix $\mathbf{U}\in\mathbb{R}^{n\times d}$ and a random vector $\mathbf{x}\in\mathbb{R}^n$ with each element i.i.d. sampled from a sub-Gaussian distribution of parameter σ_0, then for $t\ge 2$,*

$$\Pr(\|\mathbf{x}^T\mathbf{U}\| \ge t\sigma_0\sqrt{d}) \le C\exp(-Ct).$$

Proof This lemma follows from the McDiarmid's inequality (Maurer and Pontil 2021, Theorem 3). In particular, we have that

$$\text{E}\|\mathbf{x}^T\mathbf{U}\| \le \sqrt{\text{E}[\mathbf{x}^T\mathbf{U}\mathbf{U}^T\mathbf{x}]} = \sqrt{\text{E}\left[\sum_{i=1}^n\mathbf{x}_i^2\sum_{j=1}^d\mathbf{U}_{ij}^2\right]} \le \sigma_0^2 d,$$

and let $\mathbf{x}^{(i)}\in\mathbb{R}^n$ be defined such that $\mathbf{x}_j^{(i)} = \mathbf{x}_j$ if $j\ne i$ and $\mathbf{x}_i^{(i)} = 0$, then $|\|\mathbf{x}^T\mathbf{U}\|-\|\mathbf{x}^{(i)T}\mathbf{U}\|| \le |\mathbf{x}_i|\|\mathbf{U}(i,:)\|$, where $\|\mathbf{U}(i,:)\|$ represents the norm of the i-th row of $\mathbf{U}$. As a result, $\|\mathbf{x}^T\mathbf{U}\|-\|\mathbf{x}^{(i)T}\mathbf{U}\|$ is sub-Gaussian with parameter $\sigma_0\|\mathbf{U}(i,:)\|$. Combining it with the fact that $\sum_{i=1}^n\|\mathbf{U}(i,:)\|^2 = d$ and the sub-Gaussian version of the McDiarmid's inequality (Maurer and Pontil 2021, Theorem 3), the lemma is proved. □

Assumption 1 and Lemma 7 imply that with a probability of at least $1-C\exp(-Cn)-C\exp(-Ct(r(q+m)+p))$,

$$\begin{aligned}\left\|\sum_{i=1}^n w_{1,i}\text{vec}\left(\Pi_{T_{(\mathbf{C}^*,\gamma^*)}(\mathcal{M})}\nabla f(\mathbf{C}^*,\gamma^*)\right)\right\| &\le tC_1t\sigma_0\sqrt{n(r(q+m)+p)},\\ \left\|\sum_{i=1}^n w_{1,i}\text{vec}\left(\Pi_{T_{(\mathbf{C}^*,\gamma^*)}(\mathcal{M}),\perp}\nabla f(\mathbf{C}^*,\gamma^*)\right)\right\| &\le tC_1t\sigma_0\sqrt{n(qm-r(q+m)+p)},\end{aligned} \tag{A5}$$

where

$$\sigma_0 = \begin{cases} \sigma, & \text{for the matrix-variate regression model;} \\ 1, & \text{for the logistic matrix-variate regression model.} \end{cases}$$

For the Hessian matrix, it is as defined in (A1). As a result, we have $C_{H,1} = C_2 n$. Plug in Lemma 5, we have that for $b = \sqrt{\|\hat{\mathbf{C}} - \mathbf{C}^*\|_F^2 + \|\hat{\gamma} - \gamma^*\|^2}$, we have

$$f(\mathbf{C}, \gamma) - f(\mathbf{C}^*, \gamma^*) \geq \frac{1}{2} b^2 C_2 - b\|\Pi_{T_{\mathbf{x}^*}(\mathcal{M})} \nabla f(\mathbf{x}^*)\| - C_T b^2 \|\Pi_{T_{\mathbf{x}^*}(\mathcal{M}),\perp} \nabla f(\mathbf{x}^*)\|$$

which is larger than λP if

$$\frac{b^2 C_2}{6} \geq \max\left(\lambda P, b\|\Pi_{T_{\mathbf{x}^*}(\mathcal{M})} \nabla f(\mathbf{x}^*)\|, C_T b^2 \|\Pi_{T_{\mathbf{x}^*}(\mathcal{M}),\perp} \nabla f(\mathbf{x}^*)\|\right),$$

or equivalently, if $C_2\sqrt{n} \geq 6C_1 t\sigma_0\sqrt{(qm+p)}$, and

$$b \geq \max\left(\frac{6C_1 t\sigma_0\sqrt{n(r(q+m)+p)}}{C_2}, \sqrt{\frac{6\lambda P(\mathbf{C}^*, \gamma^*)}{C_2}}\right). \tag{A6}$$

As a result, we have

$$f(\mathbf{C}, \gamma) - f(\mathbf{C}^*, \gamma^*) \geq \lambda P(\mathbf{C}^*, \gamma^*) \quad \text{for all } \{(\mathbf{C}, \gamma) \in \mathcal{M} : \text{dist}((\mathbf{C}, \gamma), (\mathbf{C}^*, \gamma^*)) \in \mathcal{I}\}, \tag{A7}$$

where

$$\mathcal{I} = \left[\max\left(\frac{6C_1 t\sigma_0\sqrt{n(r(q+m)+p)}}{C_2}, \sqrt{\frac{6\lambda P(\mathbf{C}^*, \gamma^*)}{C_2}}\right), c_0\right].$$

Next, for all $(\mathbf{C}, \gamma)$ such that $\text{dist}((\mathbf{C}, \gamma), (\mathbf{C}^*, \gamma^*)) = \sqrt{\|\mathbf{C} - \mathbf{C}^*\|_F^2 + \|\gamma - \gamma^*\|^2} = b$ where $b \leq c_0$, we have

$$f(\mathbf{C}, \gamma) - f(\mathbf{C}^*, \gamma^*) \geq \frac{1}{2} C_2 n b^2 - b\|\nabla f(\mathbf{C}^*, \gamma^*)\| \geq \frac{1}{2} C_2 n b^2 - b C_1 t\sigma_0\sqrt{n(qm+p)}.$$

That is, when $b \geq \frac{4C_1 t\sigma_0\sqrt{n(qm+p)}}{C_2 n}$ and $\frac{1}{4} C_2 n b^2 \geq \lambda P(\mathbf{C}^*, \gamma^*)$. By (3.1), such a choice of b exists and we have $f(\mathbf{C}, \gamma) - f(\mathbf{C}^*, \gamma^*) \geq \lambda P(\mathbf{C}^*, \gamma^*)$ for all $\{(\mathbf{C}, \gamma) : \text{dist}((\mathbf{C}, \gamma), (\mathbf{C}^*, \gamma^*)) = b\}$. Since f is convex, it holds for all $\{(\mathbf{C}, \gamma) : \text{dist}((\mathbf{C}, \gamma), (\mathbf{C}^*, \gamma^*)) \geq b\}$. Combining it with (A7), we have that for all $(\mathbf{C}, \gamma) \in \mathcal{M}$ such that $\sqrt{\|\mathbf{C} - \mathbf{C}^*\|_F^2 + \|\gamma - \gamma^*\|^2} \geq C_{error,1}$, $f(\mathbf{C}, \gamma) - f(\mathbf{C}^*, \gamma^*) \geq \lambda P(\mathbf{C}^*, \gamma^*)$, and the theorem is proved. □

A.7 Proof of Theorem 4

Proof First, by Assumption 2, for all $\{(\mathbf{C}, \gamma) : \sqrt{\|\mathbf{C} - \mathbf{C}^*\|_F^2 + \|\gamma - \gamma^*\|^2} = c_0'^2\}$,

$$F(\mathbf{C}, \gamma) \geq f(\mathbf{C}, \gamma) > \frac{C_2 n}{2} c_0'^2 + \sum_{i=1}^{n} l(y_i, \langle \mathbf{X}_i, \mathbf{C}^* \rangle + \gamma^{*T} \mathbf{z}_i.$$

Since $F(\mathbf{C}^{(\text{iter})}, \gamma^{(\text{iter})})$ is nonincreasing, we can choose a small initial step size $\alpha_0 > 0$, such that if the initial step size α in line search satisfies $\alpha < \alpha_0$, then $(\mathbf{C}^{(\text{iter})}, \gamma^{(\text{iter})}) \in \mathcal{B}$ for all iter ≥ 1.

By the proof of (A5) we have

$$\|T_{(\mathbf{C}^*,\gamma^*),\mathcal{M}} \nabla f(\mathbf{C}^*, \gamma^*)\| \leq nC_1 t\sigma_0\sqrt{n(r(q+m)+p)}$$
$$\|T_{(\mathbf{C}^*,\gamma^*),\mathcal{M},\perp} \nabla f(\mathbf{C}^*, \gamma^*)\| \leq nC_1 t\sigma_0\sqrt{n(qm - r(q+m)+p)}.$$

As a result, for $(\mathbf{C}, \gamma)$ with $\text{dist}((\mathbf{C}, \gamma), (\mathbf{C}^*, \gamma^*)) = b$, we have

$$\|T_{(\mathbf{C}^*,\gamma^*),\mathcal{M}} \nabla f(\mathbf{C}^*, \gamma^*)\| \geq nC_2 b - C_1 t\sigma_0\sqrt{n(r(q+m)+p)}$$
$$\|T_{(\mathbf{C}^*,\gamma^*),\mathcal{M},\perp} \nabla f(\mathbf{C}^*, \gamma^*)\| \leq nC_3 b + C_1 t\sigma_0\sqrt{n(qm - r(q+m)+p)}.$$

That is, if

$$nC_2 b - C_1 t\sigma_0\sqrt{n(r(q+m)+p)} > C_T b(nC_3 b + C_1 t\sigma_0\sqrt{n(qm - r(q+m)+p)}) + \lambda C_{partial},$$

then $\|T_{(\mathbf{C},\gamma),\mathcal{M}} \nabla f(\mathbf{C}^*, \gamma^*)\| \neq 0$. This is satisfied if

$$\frac{1}{4} nC_2 b > \max\left\{C_1 t\sigma_0\sqrt{n(r(q+m)+p)}, C_T C_3 n b^2, C_T C_1 t\sigma_0\sqrt{n(qm - r(q+m)+p)}b, \lambda C_{partial}\right\},$$

i.e., when $\sqrt{n} > \frac{4C_T C_1 t\sigma_0\sqrt{(qm - r(q+m)+p)}}{C_2}$ and

$$\frac{4C_1 t\sigma_0\sqrt{(r(q+m)+p)}}{C_2\sqrt{n}} + \frac{4\lambda C_{partial}}{nC_2} < b < \frac{C_2}{4C_T C_3}.$$

By assumptions, this is satisfied with initialization $b = \text{dist}((\mathbf{C}^{(0)}, \gamma^{(0)}), (\mathbf{C}^*, \gamma^*))$, so

$$(\mathbf{C}^{(\text{iter})}, \gamma^{(\text{iter})}) \in \mathcal{B} \ \forall \text{ iter} \geq 1.$$

It remains to prove (3.3), which is similar to the proof of (3.2). □

A.8 Proof of Theorem 3

Proof of Theorem 3 WLOG assume that $m \geq q$. Let $\theta = (\mathbf{C}, \gamma)$, and

$$\mathcal{C} = \Big\{(\mathbf{C}, \gamma) : \mathbf{C} = [\mathbf{C}', 0] \text{ where } \mathbf{C}' \in \mathbb{R}^{m\times r} \text{ and } 0 \in \mathbb{R}^{m\times(q-r)}, \mathbf{C}'_{ij} \in \{s, -s\}, \gamma_k \in \{s, -s\}\Big\},$$

where $s = c_0\sqrt{\frac{r(m+q)+p}{n}}$. Then we have $|\mathcal{C}| = 2^{rm+p}$, and for any $(\mathbf{C}_1, \gamma_1), (\mathbf{C}_2, \gamma_2) \in \mathcal{C}$, $\text{dist}((\mathbf{C}_1, \gamma_1), (\mathbf{C}_2, \gamma_2)) \geq 2s$. In addition, for any $(\mathbf{C}, \gamma) \in \mathcal{C}$ and P_0 represents the model when $\mathbf{C} = 0$ and $\gamma = 0$,

$$\begin{aligned} K(P_0, P_{(\mathbf{C},\gamma)}) &\leq c_\epsilon \sum_{i=1}^{n} \Big(\langle \mathbf{X}_i, \mathbf{C}\rangle + \gamma^T \mathbf{z}_i\Big)^2 \\ &\leq c_\epsilon C_{upper}(\|\mathbf{C}\|_F^2 + \|\gamma\|^2) \\ &\leq c_\epsilon C_{upper}(r(m+q)+p)s^2. \end{aligned}$$

Applying (Tsybakov 2008, Theorem 2.5) and note that $\log|\mathcal{C}| = \log(2^{rm+p}) = (rm+p)\log 2$, we may choose $\alpha = 2c_\epsilon C_{upper} s^2/\log 2$, and α can be sufficiently small by choosing c_0 to be small. The rest of the proof following applying (Tsybakov 2008, Theorem 2.5). □

References

Absil, P.A., Mahony, R., Sepulchre, R.: Optimization Algorithms on Matrix Manifolds. Princeton University Press, Princeton (2009)

Absil, P.-A., Oseledets, I.V.: Low-rank retractions: a survey and new results. Comput. Optim. Appl. **62**(1), 5–29 (2015). https://doi.org/10.1007/s10589-014-9714-4

Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J. Imag. Sci. **2**(1), 183–202 (2009)

Boumal, N.: On intrinsic cramér-rao bounds for Riemannian submanifolds and quotient manifolds. IEEE Trans. Signal Process. **61**(7), 1809–1821 (2013). https://doi.org/10.1109/TSP.2013.2242068

Boumal, N., Mishra, B., Absil, P.-A., Sepulchre, R.: Manopt, a matlab toolbox for optimization on manifolds. J. Mach. Learn. Res. **15**(1), 1455–1459 (2014)

Boumal, N., Mishra, B., Absil, P.-A., Sepulchre, R.: Manopt, a Matlab toolbox for optimization on manifolds. J. Mach. Learn. Res. **15**(42), 1455–1459 (2014)

Campbell, N.A.: Robust procedures in multivariate analysis i: Robust covariance estimation. J. Roy. Stat. Soc.: Ser. C (Appl. Stat.) **29**(3), 231–237 (1980)

Candès, E.J., Recht, B.: Exact matrix completion via convex optimization. Found. Comput. Math. **9**(6), 717–772 (2009)

Candes, E.J., Tao, T.: Decoding by linear programming. IEEE Trans. Inf. Theory **51**(12), 4203–4215 (2005)

Caruana, R.: Multitask learning. Mach. Learn. **28**(1), 41–75 (1997)

Chen, H., Guo, Y., He, Y., Ji, J., Liu, L., Shi, Y., Wang, Y., Yu, L., Zhang, X., Initiative, A.D.N., et al.: Simultaneous differential network analysis and classification for matrix-variate data with application to brain connectivity. Biostatistics **23**(3), 967–89 (2021)

Choi, Y., Taylor, J., Tibshirani, R.: Selecting the number of principal components: estimation of the true rank of a noisy matrix. Annals Stat. **1**, 2590–2617 (2017)

Donoho, D.L.: Compressed sensing. IEEE Trans. Inf. Theory **52**(4), 1289–1306 (2006)

Elsener, A., Geer, S.: Robust low-rank matrix estimation. Ann. Stat. **46**(6B), 3481–3509 (2018)

Epstein, C.: American clinical neurophysiology society guideline 5: Guidelines for standard electrode position nomenclature. J. Clin. Neurophysiol. **23**(2), 107–110 (2006)

Fan, J., Wang, W., Zhu, Z.: A shrinkage principle for heavy-tailed data: high-dimensional robust low-rank matrix recovery. Ann. Stat. **49**(3), 1239–1266 (2021). https://doi.org/10.1214/20-AOS1980

Hao, R., Wang, X., Zhang, J., Liu, J., Du, X., Liu, L.: Automatic detection of fungi in microscopic leucorrhea images based on convolutional neural network and morphological method. In: 2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), IEEE, pp. 2491–2494, (2019)

Huang, H.-H., Zhang, T.: Robust discriminant analysis using multi-directional projection pursuit. Pattern Recogn. Lett. **138**, 651–656 (2020)

Huber, P.J.: Robust estimation of a location parameter. Ann. Math. Stat. **35**(1), 73–101 (1964). https://doi.org/10.1214/aoms/1177703732

Hung, H., Jou, Z.-Y.: A low rank-based estimation-testing procedure for matrix-covariate regression. Stat. Sin. **29**(2), 1025–1046 (2019)

Hung, H., Wang, C.-C.: Matrix variate logistic regression model with application to EEG data. Biostatistics **14**(1), 189–202 (2012). https://doi.org/10.1093/biostatistics/kxs023

Koltchinskii, V., Lounici, K., Tsybakov, A.B.: Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. Ann. Stat. **39**(5), 2302–2329 (2011)

Le Cam, L.: Maximum likelihood: an introduction. Int. Stat. Rev. **1**, 153–171 (1990)

Li, M., Kong, L., Su, Z.: Double fused lasso regularized regression with both matrix and vector valued predictors. Electron. J. Stat. **15**(1), 1909–1950 (2021)

Lu, Z., Monteiro, R.D., Yuan, M.: Convex optimization methods for dimension reduction and coefficient estimation in multivariate linear regression. Math. Program. **131**(1–2), 163–194 (2012)

Luo, Y., Huang, W., Li, X., Zhang, A.R.: Recursive importance sketching for rank constrained least squares: algorithms and high-order convergence. arXiv preprint arXiv:2011.08360 (2020)

Maronna, R.A., Martin, R.D., Yohai, V.J., Salibián-Barrera, M.: Robust Statistics: Theory and Methods (with R). Wiley Series in Probability and Statistics. Wiley, Armstrong (2018)

Maurer, A., Pontil, M.: Concentration inequalities under sub-gaussian and sub-exponential conditions. In: Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems (2021). https://openreview.net/forum?id=WJPAqX5M-2

Negahban, S., Wainwright, M.J.: Estimation of (near) low-rank matrices with noise and high-dimensional scaling. Ann. Stat. **39**(2), 1069–1097 (2011). https://doi.org/10.1214/10-AOS850

Recht, B., Fazel, M., Parrilo, P.A.: Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. SIAM Rev. **52**(3), 471–501 (2010). https://doi.org/10.1137/070697835

Rohde, A., Tsybakov, A.B.: Estimation of high-dimensional low-rank matrices. Ann. Stat. **39**(2), 887–930 (2011). https://doi.org/10.1214/10-AOS860

She, Y., Chen, K.: Robust reduced-rank regression. Biometrika **104**(3), 633–647 (2017)

She, Y., Wang, Z., Jin, J.: Analysis of generalized Bregman surrogate algorithms for nonsmooth nonconvex statistical learning. Ann. Stat. **49**(6), 3434–3459 (2021)
Tibshirani, R.: Regression shrinkage and selection via the lasso. J. Roy. Stat. Soc.: Ser. B (Methodol.) **58**(1), 267–288 (1996)
Tsybakov, A.B.: Introduction to Nonparametric Estimation. Springer Series in Statistics. Springer, USA (2008)
Vandereycken, B.: Low-rank matrix completion by Riemannian optimization. SIAM J. Optim. **23**(2), 1214–1236 (2013)
Wainwright, M.J.: High-Dimensional Statistics: A Non-Asymptotic Viewpoint. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge (2019). https://doi.org/10.1017/9781108627771
Wang, X., Zhu, H., Initiative, A.D.N.: Generalized scalar-on-image regression models via total variation. J. Am. Stat. Assoc. **112**(519), 1156–1168 (2017)
Zhang, T., Yang, Y.: Robust PCA by manifold optimization. J. Mach. Learn. Res. **19**(1), 3101–3139 (2018)
Zhang, X.L., Begleiter, H., Porjesz, B., Wang, W., Litke, A.: Event related potentials during object recognition tasks. Brain Res. Bull. **38**(6), 531–538 (1995)
Zhou, H., Li, L.: Regularized matrix regression. J. Roy. Stat. Soc.: Ser. B (Stat. Methodol.) **76**(2), 463–483 (2014)