

Nonparametric learning with matrix-valued predictors in high dimensions

Abstract

We consider the problem of learning the relationship between a binary label response and a high-dimensional matrix-valued predictor. Prediction based on matrices or networks has recently surged in brain connectivity studies, sensor network localization, and integrative genomics. Traditional regression methods take a parametric procedure by imposing a priori functional form between variables. These parametric models, however, are inadequate for structure learning and often fail in accurate prediction. Here, we develop a learning reduction framework to address a range of learning tasks from classification to regression in the matrix-valued predictor space. Our proposal achieves interpretable prediction via a low-rank two-way sparse representation of the target function. Unlike earlier approaches, our method automatically learns and exploits the important features in the high-dimensional matrices. Statistical accuracy, excess risk bounds, and efficient algorithms are established. We demonstrate the advantage of our method over previous approaches through simulations and applications to human brain connectome data.

Keywords: Nonparametric learning, high-dimensional matrices, sparse and low-rank models, classification, regression, feature selection

1 Introduction

Matrix-valued predictors ubiquitously arise in modern applications. In brain connectivity studies, for example, each individual is represented by their own brain network, and the network quantifies the connectivity pattern between a common set of nodes (brain regions of interest). Human connectome project ([Wang et al., 2017](#)) constructs the brain networks using Desikan atlas with 68 nodes over multiple cortical regions. Structural connectivity is measured for every pair of nodes, resulting in an adjacency matrix of size 68×68 for each individual. This connectivity matrix provide important information for disease prediction. Other examples include electroencephalography studies of alcoholism ([Zhou and Li, 2014](#)). Researchers measure the voltage values from 64 channels of electrodes on 256 subjects for 256 time points. The study yields a 256×64 matrix-valued feature, along with a binary indicator of subject being alcoholic or not. Identifying the relationship between

EEG signals and alcoholism is helpful for disease diagnostics.

Related work. We consider the statistical learning problem to model the relationship between a matrix-valued predictor $\mathbf{X} \in \mathbb{R}^{d_1 \times d_2}$ and a binary label response $y \in \{-1, 1\}$. The key challenge with matrix-valued predictors is the high-dimensional multi-way structure in the feature space. One approach is to transform the predictors into vectors and apply classical methods such as Lasso (Friedman et al., 2010). The practice of vectorization, however, destroys the structural information in the original predictors. Indeed, network data encoded as matrices represent various aspects of features, including global structure (e.g. clustering patterns, community hubs) and local structure (e.g. node degrees, edge connections). Learning and incorporating these features are important for prediction. There have been several recent attempts to allow matrix-valued predictors; for example, trace regression (Fan et al., 2019), network logistic regression (Reli3n et al., 2019), and matrix linear discriminant analysis (Hu et al., 2020). These parametric approaches impose a priori functional form between variables and often lead to inaccurate prediction in high dimensions. For these reasons, nonparametric approaches such as k -nearest neighbors, decision trees, and convolutional neural network (CNN) have been popular. Current nonparametric methods sacrifice interpretability to trade for prediction accuracy. In our motivating brain network application and many other scientific studies, however, researchers are interested in *interpretable prediction*, where the goal is to not only make accurate prediction but also identify most informative features for descriptive simplicity. Efficient methods for matrix based prediction have yet to be developed.

Our contributions. We develop a nonparametric learning framework that automatically exploits the matrix-valued features for accurate prediction. We address three matrix problems – classification, level set estimation, and regression – via a learning reduction approach. The proposal achieves interpretable prediction using a low-rank two-way sparse representation of the target functions. We establish theoretical accurate guarantees that permits the matrix dimension to greatly exceed the sample size. Unlike earlier approaches, our method performs efficient variable selection and adapts to the possibly non-smooth, non-linear functions of interest. Our numerical analyses and application demonstrate the outperformance of the proposed approach over previous methods.

Notation. Let $\mathcal{X} = \mathbb{R}^{d_1 \times d_2}$ be the feature space. Given a function $f: \mathcal{X} \rightarrow \mathbb{R}$, we use $\text{sign}f$ to denote its sign function such that $\text{sign}f(\mathbf{X}) = 1$ if $f(\mathbf{X}) > 0$ and $\text{sign}f(\mathbf{X}) = -1$ otherwise. The

notion of sign function also extends to sets in \mathcal{X} . We use $\text{sign}(\mathbf{X} \in A)$ to denote the sign function induced by the set $A \subset \mathcal{X}$, i.e., a function taking value 1 on the event $\{\mathbf{X} \in A\}$ and -1 otherwise. We use $\mathbf{1}(\cdot)$ to denote the indicator function, the shorthand $[n] := \{1, \dots, n\}$ to denote the n -set for $n \in \mathbb{N}_+$, and $|\cdot|$ to denote the cardinality of sets. Let $\|\cdot\|_p$ denote the vector p -norm for $p \geq 0$, and $\|\cdot\|_F$ be the matrix Frobenious norm. The (p, q) -norm of a matrix \mathbf{B} is defined as $\|\mathbf{B}\|_{p,q} = \|\mathbf{b}\|_q$, where $\mathbf{b} = (\|\mathbf{B}_1\|_p, \dots, \|\mathbf{B}_{d_1}\|_p)^T \in \mathbb{R}^{d_1}$ consists of the p -norms for each of the rows in \mathbf{B} , and \mathbf{B}_i is the i -th row of matrix \mathbf{B} for $i \in [d_1]$. In particular, $\|\mathbf{B}\|_{1,0} = |\{i \in [d_1]: \mathbf{B}_i \neq 0\}|$ denotes the number of non-zero rows in \mathbf{B} . An event E is said to occur “with high probability” if $\mathbb{P}(E)$ tends to 1 as the matrix dimension $d_{\min} = \min(d_1, d_2) \rightarrow \infty$. We denote $a_n \asymp b_n$ if $\lim_n b_n/a_n \rightarrow c$ for some constant $c > 0$ and denote $a_n \lesssim b_n$ if $\lim_n b_n/a_n \rightarrow 0$.

2 Three learning problems

We present the main learning goals of our interest. Let $\mathbf{X} \in \mathbb{R}^{d_1 \times d_2}$ denote the matrix-valued predictor, $y \in \{-1, 1\}$ denote the binary label response, and $\mathbb{P}_{\mathbf{X},y}$ denote the unknown joint probability distribution over the pair (\mathbf{X}, y) . In the context of binary response, y is a Bernoulli random variable with conditional probability $p(\mathbf{X}) \stackrel{\text{def}}{=} \mathbb{P}(y = 1|\mathbf{X})$; we generally make no parametric assumptions on the marginal distribution $\mathbb{P}_{\mathbf{X}}$ or form of $p(\mathbf{X})$.

Suppose that we observe a sample of n training data points, $(\mathbf{X}_1, y_1), \dots, (\mathbf{X}_n, y_n)$, identically and independently distributed (i.i.d.) according to $\mathbb{P}_{\mathbf{X},y}$. Let $(\mathbf{X}_{\text{new}}, y_{\text{new}})$ be a future unseen test point drawn independently from the same distribution. Our goal is to predict y_{new} based on \mathbf{X}_{new} . We often omit the subscript “new” and simply write (\mathbf{X}, y) for the prototypical test point. The relevant probabilistic statements should be interpreted as taken jointly with respect to (\mathbf{X}, y) .

We consider three learning problems: classification, level set estimation, and regression.

2.1 Matrix classification: Classification is the problem of predicting the label $y \in \{-1, 1\}$ to which the new matrix \mathbf{X} belongs. A prediction rule (also called a classifier) decides that $y = 1$ if $\mathbf{X} \in S$ and $y = 0$ if $\mathbf{X} \notin S$, where S is a Borel subset of $\mathbb{R}^{d_1 \times d_2}$. We formulate the classification problem as choosing a classifier $S \in \mathcal{S}$, from a given set of candidate classifiers \mathcal{S} , that minimizes the expected classification error (also called the classification risk),

$$R(S) = \mathbb{P}_{\mathbf{X},y} [y \neq \text{sign}(\mathbf{X} \in S)]. \quad (1)$$

When the population distribution $\mathbb{P}_{\mathbf{X},y}$ is known, the global minimizer of $R(S)$ over all Borel sets of $\mathbb{R}^{d_1 \times d_2}$ is called Bayes classifier, expressed as $S_{\text{bayes}} = \{\mathbf{X} \in \mathbb{R}^{d_1 \times d_2} : p(\mathbf{X}) \geq 1/2\}$. In practice the population distribution is unknown, so the objective function (1) and the minimizer needs to be estimated through the data $\{\mathbf{X}_i, y_i\}_{i \in [n]}$. Our first goal is matrix classification:

Question 1. How to perform classification when matrix dimension far exceeds the sample size n ?

2.2 Level set estimation: The problem of level set estimation generalizes the classification task. For a given $\pi \in (0, 1)$, the target π -level set of the conditional probability function $p(\mathbf{X})$ is defined as

$$S_{\text{bayes}}(\pi) = \{\mathbf{X} \in \mathbb{R}^{d_1 \times d_2} : p(\mathbf{X}) \geq \pi\}.$$

An important fact is that the set $S_{\text{bayes}}(\pi)$ optimizes the weighted classifier risk (Willett and Nowak, 2007; Scott and Davenport, 2007; Wang et al., 2008). Specifically, among all Borel subsets of $\mathbb{R}^{d_1 \times d_2}$, the set $S_{\text{bayes}}(\pi)$ is the global minimizer of the expected π -weighted classification error,

$$R_\pi(S) = \mathbb{E} [w_\pi(y) \mathbb{1}(y \neq \text{sign}(\mathbf{X} \in S))], \quad (2)$$

where we define $w_\pi(y) = 1 - \pi$ or π depending on $y = 1$ or -1 . In light of (1) and (2), the level set estimation is an extension of the usual classification from equal weight $\pi = 1/2$ to general weight $\pi \in (0, 1)$. Accurate level set estimation plays an important role in applications of geographical elevation maps, imaging contour detection, and motion tracking. We consider the following question:

Question 2. How to simultaneously estimate the level set and identify important variables in the matrix-valued predictor space, for the goal of interpretable prediction?

2.3 Nonparametric regression: The problem of nonparametric regression is to estimate the conditional mean $\mathbb{E}(y|\mathbf{X})$ as a multivariable function in the predictor space. In the our contexts, the nonparametric regression is equivalent to estimating the conditional probability $p(\mathbf{X}) = \mathbb{P}(y = 1|\mathbf{X}) = \frac{1}{2}(\mathbb{E}(y|\mathbf{X}) + 1)$. Throughout the paper we will focus on $p(\mathbf{X})$ and refer to it as the regression function. Our final goal is the function estimation:

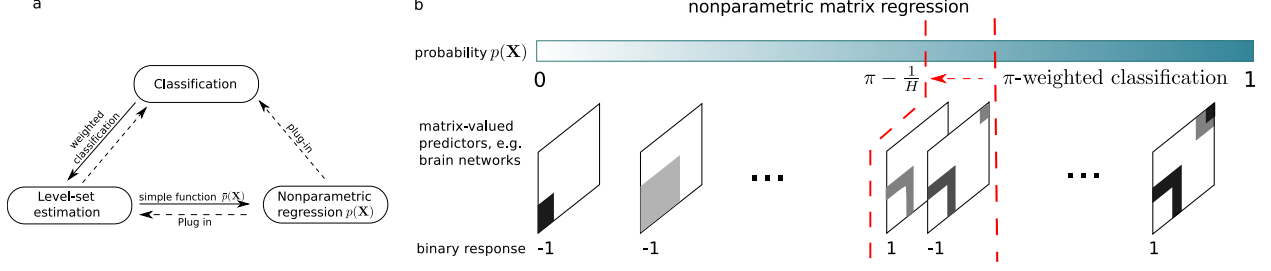


Figure 1: (a) Our learning reduction approach (solid line) to three learning problems, in contrast to the classical plug-in approaches (dashed line). (b) Matrix nonparametric regression via π -weighted classification.

Question 3. How to learn the regression function $p(\mathbf{X})$ in the high-dimensional matrix space?

The three problems of our interest represent a range of learning tasks with increasing difficulties. Classification is a special case of level set estimation with $\pi = 1/2$, whereas the level set is a discrete approximation of the regression function. A common approach is to address regression first, and then solve the earlier two using plug-in estimates (Figure 1a). This procedure, however, undermines the fact that regression is generally harder than the other two. Indeed, as we show in Section 4, regression has a slower convergence rate $\mathcal{O}(n^{-1/2})$ compared to the fast rate $\mathcal{O}(n^{-1})$ of classification. Ignorance of the increased complexity violates Vapnik’s maxim: *When solving a given problem, one should try to avoid solving a more general problem as an intermediate step.*

3 From classification to regression: a new deal

We develop a “learning reduction” approach by relating the regression to classification, the latter of which is more fundamental and easier to address. We addresses classification first and use the results to solve the regression based on learning reduction framework (Figure 1b). In general, regression requires more assumptions than classification. Our learning reduction approach bridges these two tasks using level set estimation, a problem lies somewhere in between. The connection allows us to disentangle complexity and leverage existing algorithms.

Figure 1b illustrates the main idea of our approaches. We use a sequence of weighted classifications to find the level sets in the matrix space, and then estimate the regression function $p(\mathbf{X}) = \mathbb{E}(y = 1|\mathbf{X})$ via level set aggregation. The level set approach bridges the two sides of a same coin – characteristic (set indicator) functions in functional analysis and weighted classifications in statistical learning. Specifically, let $\Pi = \{\frac{1}{H}, \frac{2}{H}, \dots, \frac{H-1}{H}\}$ be a sequence of evenly spaced points in $[0, 1]$, where $H \in \mathbb{N}_+$ is the resolution parameter. We propose an H -step function estimate

$\hat{p}(\cdot): \mathbb{R}^{d_1 \times d_2} \rightarrow [0, 1]$ by

$$\hat{p}(\mathbf{X}) = \frac{1}{2H} \sum_{\pi \in \Pi} \text{sign}(\mathbf{X} \in \hat{S}(\pi)) + \frac{1}{2}, \quad \text{for all } \mathbf{X} \in \mathbb{R}^{d_1 \times d_2}, \quad (3)$$

where, for every $\pi \in \Pi$, the set $\hat{S}(\pi) \subset \mathbb{R}^{d_1 \times d_2}$ is the estimated classifier from empirical surrogate risk minimization,

$$\hat{S}(\pi) = \{\mathbf{X}: \hat{f}_\pi(\mathbf{X}) \geq 0\}, \quad \text{with} \quad \hat{f}_\pi = \min_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^n w_\pi(y_i) \ell(y_i f(\mathbf{X}_i)) + \lambda \|f\|_F^2 \right\}, \quad (4)$$

where $w_\pi(y) = 1 - \pi$ if $y = 1$ and $w_\pi(y) = \pi$ if $y = -1$ is the label-dependent weight; $\ell(z): \mathbb{R} \mapsto \mathbb{R}_{\geq 0}$ is the surrogate classification loss defined as a function of margin $z = yf(\mathbf{X})$; $\lambda > 0$ is the penalty parameter; \mathcal{F} is a candidate function family and $\|f\|_F$ denotes the penalization, both of which will be detailed in the next paragraph. The empirical risk minimization (4) is a sample version of the population formulation (2), where we have replaced the binary loss by a more manageable large-margin loss. Examples of large-margin loss are hinge loss $\ell(z) = (1 - z)_+$ for support vector machines, logistic loss $\ell(z) = \log(1 + e^{-z})$ for important vector machines, and ψ -loss $\ell(z) = 2 \min(1, (1 - z)_+)$ with $z_+ = \max(z, 0)$. We choose hinge loss for parsimony; our framework applies equally to common large-margin losses (Bartlett et al., 2006).

We now describe choice of \mathcal{F} in (4). A desirable \mathcal{F} should balance the prediction and interpretability; i.e., \mathcal{F} should be flexible enough for accurate prediction while being simple enough for high interpretability. We propose the function family \mathcal{F} to be linear classifiers with low-rank two-way sparse matrix coefficients,

$$\mathcal{F}(r, s_1, s_2) = \{f: \mathbf{X} \mapsto \langle \mathbf{X}, \mathbf{B} \rangle + b \mid \text{rank}(\mathbf{B}) \leq r, \text{supp}(\mathbf{B}) \leq (s_1, s_2), \mathbf{B} \in \mathbb{R}^{d_1 \times d_2}, b \in \mathbb{R}\}, \quad (5)$$

where $\text{rank}(\mathbf{B})$ denotes the rank of the coefficient matrix, and $\text{supp}(\mathbf{B})$ denotes the two-way sparsity, with $s_1 = \|\mathbf{B}\|_{1,0}$ and $s_2 = \|\mathbf{B}^T\|_{1,0}$ being the numbers of non-zero rows and columns of \mathbf{B} , respectively. Correspondingly, we define the penalization term in (4) as $\|f\|_F = \|\mathbf{B}\|_F$.

The low-rank two-way sparse classifier (5) enables efficient variable selection for high-dimensional matrices, thereby achieving high interpretability in prediction. In the brain network analysis, scientists are interested in identifying important nodes attached to at least one active edges with non-zero effects. Classical entrywise sparsity essentially treats \mathbf{X} as “a bag of non-ordered edges”, and

loses the two-way paring information among entries. In contrast, our two-way sparsity efficiently identifies the underlying active nodes by making use of matrix structure in the predictors.

Many nonlinear functions in existing literature are special cases of our representation, in the sense that the true regression function $p(\mathbf{X})$ can be precisely recovered by the candidate classifiers in (5). We provide two examples here, although more general formulations are also possible.

Example 1 (Single index models (Alquier and Biau, 2013; Ganti et al., 2017)). Suppose the true regression function can be expressed as $p(\mathbf{X}) = g(\langle \mathbf{X}, \mathbf{B} \rangle)$, where $g(\cdot): \mathbb{R} \rightarrow [0, 1]$ is an arbitrary monotonic function, and \mathbf{B} is a low-rank two-way sparse matrix. Then, for every $\pi \in (0, 1)$, $\text{sign}(p(\mathbf{X}) - \pi) = \text{sign}f(\mathbf{X})$ for some $f \in \mathcal{F}(r, s_1, s_2)$. Our method generalizes single index model to high dimensional matrices by joint learning matrix coefficient \mathbf{B} and nonlinear function g .

Example 2 (Multivariate normal mixtures (Hu et al., 2020)). Suppose the matrix-valued predictor \mathbf{X} follows a Gaussian mixture distribution, $\mathbf{X}|\{y = -1\} = \mathbf{B}_1 + \mathbf{E}_1$ and $\mathbf{X}|\{y = 1\} = \mathbf{B}_2 + \mathbf{E}_2$, where $(\mathbf{B}_1 - \mathbf{B}_2)$ is a low-rank two-way sparse matrix, and $\mathbf{E}_1, \mathbf{E}_2$ are two mutually independent noise matrices with i.i.d. $N(0, 1)$ entries. Then, for every $\pi \in (0, 1)$, $\text{sign}(p(\mathbf{X}) - \pi) = \text{sign}f(\mathbf{X})$ for some $f \in \mathcal{F}(r, s_1, s_2)$. More generally, we have established the characterization by extending two classes of \mathbf{X} to a series of $\mathbf{X} = \mathbf{X}(\pi)$ over a continuous spectrum of $\pi \in (0, 1)$ (results not shown).

Combining formulations (3), (4) and (5) yields our “learning deduction” approach to nonparametric matrix regression. We conclude this section by providing an efficient alternating direction method of multipliers (ADMM) approach for solving the optimization. The procedure is summarized in Algorithm 1. The software for both matrix classification and regression is available at CRAN.

Algorithm 1: Matrix classification and level-set estimation (ADMM)

Input: Data $\{(\mathbf{X}_i, y_i) \in \mathbb{R}^{d_1 \times d_2} \times \{-1, 1\} : i \in [n]\}$, rank r , support (s_1, s_2) , ridge parameter λ , the target level $\pi \in \Pi$.

Initialize: primal variable \mathbf{B} , dual variable \mathbf{S} , Lagrangian multiplier $\mathbf{\Lambda} = \mathbf{0}$, step size $\rho = 0.1$.

Maximize $L(\mathbf{B}, \mathbf{S}, \mathbf{\Lambda}, \rho) = \frac{1}{n} \sum_{i=1}^n w_\pi(y_i) \ell(y_i \langle \mathbf{X}_i, \mathbf{B} \rangle) + \lambda \|\mathbf{B}\|_F^2 + \rho \|\mathbf{B} - \mathbf{S}\|_F^2 + \langle \mathbf{\Lambda}, \mathbf{B} - \mathbf{S} \rangle$

Do until converges

Update \mathbf{B} fixing $(\mathbf{S}, \mathbf{\Lambda}, \rho)$: Solve $L(\mathbf{B}|\mathbf{S}, \mathbf{\Lambda}, \rho) = \frac{1}{n} \sum_{i=1}^n w_\pi(y_i) \ell(y_i \langle \mathbf{X}_i, \mathbf{B} \rangle) + (\lambda + \rho) \|\mathbf{B} - \bar{\mathbf{S}}\|_F^2$.
 $\mathbf{B} \leftarrow \arg \min_{\mathbf{B}} L(\mathbf{B}|\mathbf{S}, \mathbf{\Lambda}, \rho)$.

Update \mathbf{S} fixing $(\mathbf{B}, \mathbf{\Lambda}, \rho)$: Solve $L(\mathbf{S}|\mathbf{B}, \mathbf{\Lambda}, \rho) = \|\mathbf{S} - \bar{\mathbf{B}}\|_F^2$, subject to $\mathbf{S} \in \mathcal{F}(r, s_1, s_2)$.
 $\mathbf{S} \leftarrow \arg \min_{\mathbf{S} \in \mathcal{F}(r, s_1, s_2)} L(\mathbf{S}|\mathbf{B}, \mathbf{\Lambda}, \rho)$.

Update $\mathbf{\Lambda}$ $\leftarrow \mathbf{\Lambda} + 2\rho(\mathbf{B} - \mathbf{S})$ and $\rho \leftarrow 1.1\rho$.

Output: Estimated π -level set $\hat{S}(\pi) = \{\mathbf{X} \in \mathbb{R}^{d_1 \times d_2} : \hat{f}(\mathbf{X}) \geq 0\}$.

4 Statistical properties

In this section we provide accuracy guarantees for the proposed nonparametric matrix regression (Figure 1b). We establish the excess risk bound for the weighted classification (4) and for the matrix regression (3). Our learning reduction approach successfully bridges these tasks based on Vapnik's maxim and achieve theoretical guarantee for each problem.

We introduce the following notion to characterize the behavior of the regression function near the level set boundaries $\partial S_{\text{bayes}}(\pi) = \{p(\mathbf{X}) = \pi\}$. The condition quantifies the uniqueness of level sets estimation from population characterization (2) (and its sample version (4)).

Definition 1 (Global regularity). We call a level $\pi \in [0, 1]$ a mass point if the level set boundary $\partial S_{\text{bayes}}(\pi)$ has non-zero measure under $\mathbb{P}_{\mathbf{X}}$. Let $\mathcal{N} = \{\pi \in [0, 1]: \mathbb{P}[p(\mathbf{X}) = \pi] \neq 0\}$ denote the collection of mass points in $p(\mathbf{X})$. A function $p(\mathbf{X})$ is called α -globally regular with $\alpha \in [0, 1]$, if

- (i) $p(\mathbf{X})$ has finitely many mass points, i.e., $|\mathcal{N}| \leq C'$ for some constant $C' < \infty$; and
- (ii) there exists a global constant $C > 0$ such that, for all $\pi \notin \mathcal{N}$,

$$\mathbb{P}_{\mathbf{X}}(|p(\mathbf{X}) - \pi| \leq t) \leq Ct^{\alpha/(1-\alpha)}, \quad \text{for } t \in (0, \rho(\pi, \mathcal{N})), \quad (6)$$

where $\rho(\pi, \mathcal{N}) \stackrel{\text{def}}{=} \min_{\pi' \in \mathcal{N}} |\pi - \pi'|$ denotes the distance from π to the nearest mass point in \mathcal{N} . When $\mathcal{N} = \emptyset$, we define $\rho(\pi, \mathcal{N}) = 1$. The right-hand-side of (6) is interpreted as zero when $\alpha = 1$.

Definition 1 uniformly controls the behavior of $p(\mathbf{X})$ across possible π . If the condition (6) holds for a fixed π , we call the function $p(\mathbf{X})$ is (π, α) -locally regular. The exponent α quantifies the concentration of probability mass $p(\mathbf{X})$, and its value turns out to play a key role in the level set accuracy from risk minimization (4). Accurate set estimation is more difficult at levels where the function is relatively flat ($\alpha = 0$), as intuition would suggest. The best rate $\alpha = 1$ corresponds to a clear separation with no point mass at the boundary. A typical intermediate case is $\alpha = 1/2$ when $p(\mathbf{X})$ has non-degenerate first-order Taylor expansion around π .

The following assumption quantifies the capability of candidate classifiers in $\mathcal{F}(r, s_1, s_2)$ for approximating the true level sets. We let $d_1 = d_2 = d$ for simplicity.

Assumption 1 (Approximation error). Let $f_{\text{bayes}, \pi}(\mathbf{X}) = \text{sign}(p(\mathbf{X}) - \pi)$ be the Bayes classifier corresponding to the π -level set. Assume there exists a sequence of functions $f_n^* \in \mathcal{F}(r, s_1, s_2)$

for which the surrogate excess risk vanishes; i.e., $R_{\ell,\pi}(f_n^*) - R_{\ell,\pi}(f_{\text{bayes},\pi}) \leq a_n$ for some sequence $a_n \rightarrow 0$ as $n, d \rightarrow \infty$. Here $R_{\ell,\pi}(f) = \mathbb{E}[w_\pi(y)\ell(yf(\mathbf{X}))]$ denotes the population surrogate risk in analogy to the empirical version in (4). Let $J_n = \|f_n^*\|_F^2$, and we allow J_n to grow with n .

We now provide the accuracy guarantee for the level set estimation (4). We consider the high dimensional regime as both the sample size n and matrix dimension d diverge, while treating (r, s_1, s_2) as fixed constants. The result demonstrates the statistical consistency of our classifier even when the matrix dimension far exceeds the sample size n .

Theorem 4.1 (Accuracy for matrix classification). Fix a level $\pi \in (0, 1)$. Consider the problem of π -level set estimation for a (π, α) -locally regular function $p(\mathbf{X})$ with $\alpha \in [0, 1]$. Suppose Assumption 1 holds, and let \hat{f}_π be the level set estimate in (4) with penalty parameter $\lambda \asymp \left(\frac{r(s_1 + s_2) \log d}{n J_n}\right)^{1/(2-\alpha)}$. Then, with high probability, the classification excess risk is bounded by

$$R_\pi(\hat{f}_\pi) - R_\pi(f_{\text{bayes},\pi}) \lesssim \max \left\{ \left(\frac{r(s_1 + s_2) \log d}{n} \right)^{1/(2-\alpha)}, a_n \right\}, \quad (7)$$

where $R_\pi(f) = \mathbb{E}[w_\pi(y)\mathbb{1}(y \neq \text{sign} f(\mathbf{X}))]$ denotes the weighted classification risk. Notice that the usual classification corresponds to $\pi = 1/2$.

Theorem 4.1 reveals the weak dependence on matrix dimension of our estimates. Consider the case when the statistical error (first term) dominates the approximation error (second term). Then, the bound (7) immediately implies the classification consistency in the high dimensional regime $d, n \rightarrow \infty$, as long as the matrix dimension d grows sub-exponentially in sample size n ; i.e., $d = o(e^n)$. This sample complexity shows the advantage of proposed low-rank two-way sparse structural models. Furthermore, we find that classification (7) reaches a fast rate $1/n$ when $\alpha = 1$, and in general the risk has rate no slower than $1/\sqrt{n}$.

We now reach the main results in this section for our nonparametric matrix regression.

Theorem 4.2 (Accuracy for nonparametric matrix regression). Let $p(\mathbf{X})$ be a α -globally regular function with $\alpha \in [0, 1]$. Consider the same setup as in Theorem 4.1. Furthermore, assume Assumption 1 holds for all $\pi \in \Pi \setminus \mathcal{N}$. Then with high probability, the estimate (3) is bounded by

$$\mathbb{E}|\hat{p}(\mathbf{X}) - p(\mathbf{X})| \lesssim \underbrace{\left(\frac{r(s_1 + s_2) \log d}{n} \right)^{\frac{\alpha}{2-\alpha}} + H \left(\frac{r(s_1 + s_2) \log d}{n} \right)^{\frac{1}{2-\alpha}}}_{\text{statistical error}} + \underbrace{a_n^\alpha}_{\text{approximation error}} + \underbrace{\frac{1}{H}}_{\text{reduction error}}.$$

Theorem 4.2 demonstrates the high dimensional convergence of our nonparametric matrix regression. Our results reveal three sources of errors: the statistical error in classification due to finite sample size, the approximation error due to the capability of classifier space $\mathcal{F}(r, s_1, s_2)$, and an additional approximation error due to learning reduction from classification to regression. The resolution parameter H controls the balances the bias-variance tradeoff.

Corollary 4.1 (High-dimensional consistency). Consider the same set-up as in Theorem 4.2. Assume $a_n \lesssim \left(\frac{r(s_1+s_2)\log d}{n}\right)^{1/(2-\alpha)}$ and set $H \asymp \left(\frac{n}{r(s_1+s_2)\log d}\right)^{1/(4-2\alpha)}$. Then, with high probability,

$$\mathbb{E}|\hat{p}(\mathbf{X}) - p(\mathbf{X})| \lesssim \left(\frac{r(s_1 + s_2)\log d}{n}\right)^{\min(1/2, \alpha)/(2-\alpha)}, \quad \text{as } d, n \rightarrow \infty \text{ while } d = o(e^n), \quad (8)$$

We conclude this section by comparing regression and classification. The fast regression error (8) is $1/\sqrt{n}$ when $\alpha = 1$. We find that the regression error (8) is generally slower than the corresponding classification rate in (7). This confirms that classification is an easier problem with faster statistical convergence. Our learning reduction approach successfully bridges these tasks based on Vapnik's maxim and achieve theoretical guarantee for each problem. In principle, more complicated classifiers, such as neural network, decision trees, and boosting, can also be brought to bear on the level set construction (4). The ability to import and adapt existing classification methods is one advantage of the proposed learning reduction framework. We expect this general principle may also benefit other settings beyond matrix learning tasks.

5 Numerical experiments

We evaluate the empirical performance of our method, and compare the accuracy with other common approaches. The simulation covers a range of nonlinear, nonsmooth models which do not necessarily follow the assumptions in our proposal. This allows us to fairly assess the performance of various approaches under practical applications.

We examine the prediction accuracy of our proposed method using multiple-index models. The multiple index model extends the single index model (see Example 1) by allowing a multi-variate latent response $(z_1, z_2) = (\langle \mathbf{B}_1, \mathbf{X} \rangle, \langle \mathbf{B}_2, \mathbf{X} \rangle)$. We simulate random matrices $\mathbf{X} \in \mathbb{R}^{d \times d}$ with i.i.d. entries from $\text{Unif}[0,1]$ and draw $\mathbf{B}_1, \mathbf{B}_2$ from $\mathcal{F}(r, s, s)$. The response is simulated by $y \sim \text{Ber}(p(\mathbf{X}))$,

where the regression function $p(\mathbf{X})$ is generated from the following chain scheme,

$$\mathbf{X} \rightarrow (z_1, z_2) \rightarrow (\bar{z}_1, \bar{z}_2) \rightarrow p(\mathbf{X}) = \begin{cases} g(\bar{z}_1), & \text{if } \bar{z}_1 > 0, \\ g(\bar{z}_2), & \text{otherwise.} \end{cases}$$

We set Φ_i = empirical cumulative distribution function (CDF) of z_i for $i = 1, 2$; Φ_0 = CDF for standard normal; $g(z) = (1 + \exp(z))^{-1}$; matrix dimension $d = 20, 30, \dots, 60$; and training sample size $n = 100, 150, \dots, 400$. The construction of p amounts to a high nonlinearity from \mathbf{X} to $p(\mathbf{X})$, and its functional form is unknown to the algorithm.

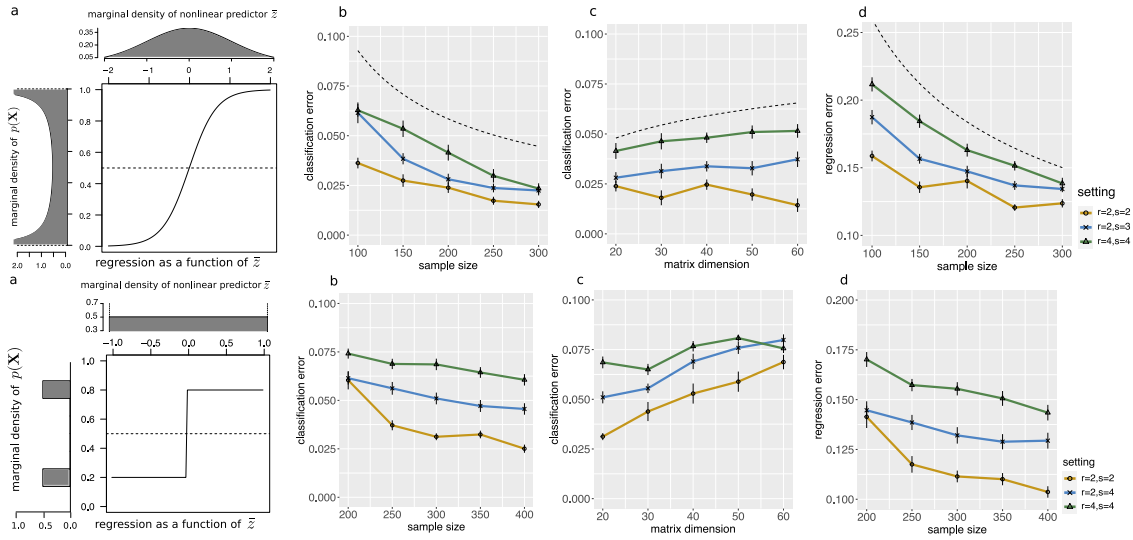


Figure 2: Prediction accuracy for matrix classification (b-c) and regression (d) under simulation setup (a).

We first assess the prediction performance under model settings depicted in Figure 2a. Figures 2b-c show the polynomial decays of classification and regression error in sample size, which is consistent with our theoretical results. We find that a higher rank or a denser coefficient leads to a higher error, as reflected by the upward shift of the curve as (r, s) increases. Indeed, a higher (r, s) implies a higher model complexity, thus increasing the generalization error. Furthermore, Figure 2d shows that error increases slowly with matrix dimension in a near-log rate. The ability to effectively control massive noisy features highlights the benefit of our method in high dimensions.

Figure 2e shows a model setting that falls on the other end of the spectrum. The random variable $p(\mathbf{X})$ concentrates at two mass points $\pi = 0.2$ and 0.8 . This makes the π -level set estimation challenging around $\pi = 0.2$ and 0.8 , because of the nonidentifiability in the weighted classification. Interestingly, we find that our method maintains good performance on classification at $\pi = 0.5$

(Figures 2f-h) and the overall regression (Figure 2g). The result demonstrate the robustness of our method to a few off-target level sets, as long as the majority are accurate.

Next, we compare our **Nonparametric MAtrix Regression (NonMAR)** with several popular alternative methods: regular lasso (**Lasso**, Friedman et al. (2010)), parametric regression for network predictors with group lasso (**LogisticM**, Relión et al. (2019)), and convolutional neural network (**CNN**). We choose a range of representative methods and aims to investigate the benefit of each approach. The **Lasso** serves as a baseline to assess the gain of matrix-valued predictors over vector-valued predictors. The methods **CNN** and **NonMAR** are nonparametric approaches and **LogisticM** is a parametric approach for matrix based prediction.

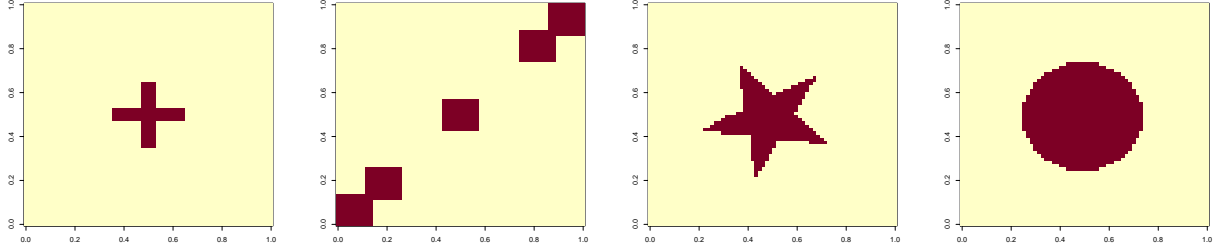


Figure 3: Active pattern models in simulations. The active region is divided into four or five subregions, each of which has its own edge connectivity signal $g_{ij}(\pi)$.

For fair comparison, we adopt similar simulation setup as in Relión et al. (2019), expecting that we add more challenging network patterns in order to assess model misspecification. We simulate observation pairs from latent variable model $(\mathbf{X}, y)|\pi$ based on the following diagram,

$$\pi \sim_{\text{i.i.d.}} \text{Unif}[0, 1] \xrightarrow{\text{conditional on } \pi} \begin{cases} y \sim \text{Ber}(\pi), \quad y \perp \mathbf{X} | \pi, \\ \mathbf{X} = \llbracket \mathbf{X}_{ij} \rrbracket, \text{ where } \mathbf{X}_{ij} \sim_{\text{i.i.d.}} \mathcal{N}(g_{ij}(\pi) \mathbf{1}(\text{edge } (i, j) \text{ is active}), \sigma^2). \end{cases}$$

The edge connectivity strength $g_{ij}(\pi)$ varies depending on the response probability π and location of $(i, j) \in [d]^2$. Figure 3 illustrate the four patterns of active regions, where each subregion has its own function $g_{ij}(\cdot): [0, 1] \rightarrow \mathbb{R}$. The function form of $g_{ij}(\cdot)$ is randomly drawn from a pre-specified library consisting of common functions such as $g(z) = \log(5z + 1), 3 \tan(z), 6z^2, \dots$. We set $d = 68$, a training size $n = 160$, and a test size 80. The hyperparameters in each method are selected by either default setting (**LogisticM**) or cross validation (**NonparaM**, **Lasso**, **CNN**).

Figure 4 compares the prediction accuracy between different methods. We find that **NonMAR**

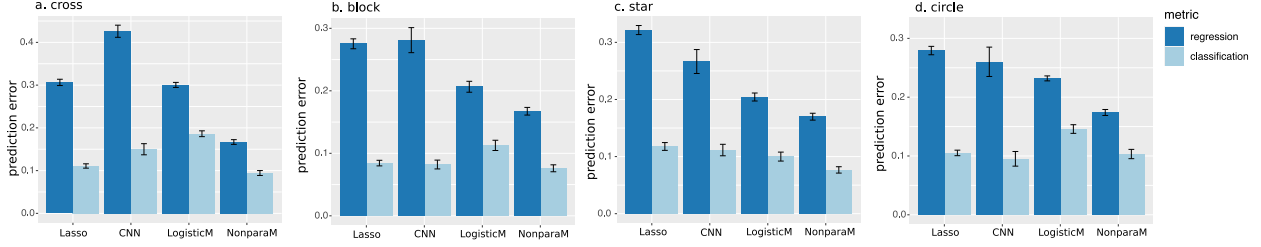


Figure 4: Performance comparison between various methods under four different active pattern models. **NonMAR** consistently outperforms others, and the reduction in regression error is substantial. For example, the relative reduction in regression error using **NonMAR** over the next best approach, **LogisticM**, is over 20% for models 1 and 4, and over 15% for models 2 and 3. In contrast, neither **Lasso** nor **CNN** has satisfactory performance. One possible reason is that these two methods fail to appropriately incorporate the network structure of the predictors. The **Lasso** takes vectorized matrices as inputs and therefore loses the two-way pairing information. On the other hand, **CNN** assumes spacial ordering within row/column indices. Although local similarity is an appropriate model for imaging, the row/column indices are meaningless for networks. The only exception is the circle model where the **CNN** has a lower classification error by a slight margin. This is perhaps due to the fact that circle pattern is nearly full rank which favors complicated models such as **CNN**. Nevertheless, our method **NonMAR** achieves stable performance in spite of its simplicity.

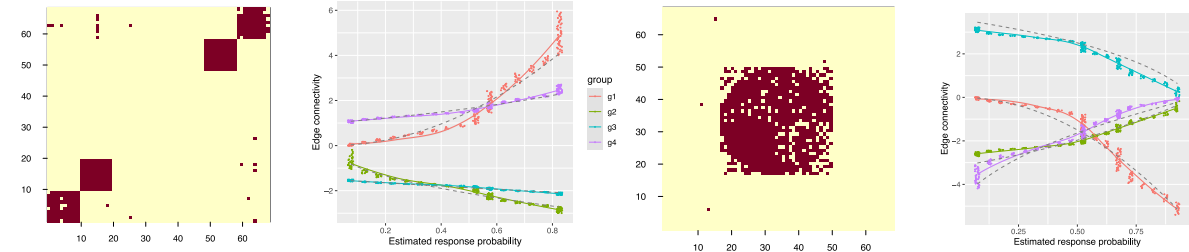


Figure 5: Example results returned by **NonMAR**. Panels (a) and (c) plot the top edges selected by **NonMAR**. Panels (b) and (d) plot the average edge connectivity for individual i against the predicted response probability $\hat{\pi}_i$. Edges are averaged by subregion, with the true function depicted in dashed curve.

We provide example results returned by **NonMAR** in typical realizations. Figures 5a and c plot the top edges selected by **NonMAR** based on the moving average of feature weights $(\hat{B}_\pi)_{\pi \in [0,1]}$ with a window size 0.2. The selected region agrees well with the ground truth (Figures 4a and c). We also investigate the edge connectivity for individual i against the predicted response probability $\hat{\pi}_i$. The trajectory of the edge connectivity accurately resembles the ground truth functions in each active subregion. The results demonstrate that our method is able to recover the right “sorting” of individuals with respect to the response probability on a continuous spectrum. This is helpful

for investigating the heterogeneous network patterns with varying disease probability.

6 Application to Human Connectome Project

We apply our method to brain network data from Human Connectome Project (HCP). The dataset is preprocessed as in Wang et al. (2017), and we analyze $n = 212$ individuals whose Variable Short Penn Line Orientation Test (VSPLLOT) scores are either high ($y = 1$) or low ($y = -1$). The VSPLLOT score measures the visuospatial processing ability. Each individual has a brain network represented by a 68-by-68 binary adjacency matrix $\mathbf{X} \in \{0, 1\}^{68 \times 68}$. The entries in the matrix encodes the presence or absence of fiber connections between the 68 brain regions. We adjust age and gender, and use a random 60-20-20 split of the data for training, validation, and testing.

Method	AUC	% of active nodes
NonMAR-p	0.73 (0.03)	89.1
NonMAR	0.77 (0.04)	97.3
LogisticM	0.72 (.02)	100
Lasso	0.68 (0.01)	27.9
CNN	0.67 (0.03)	-

Table 1: Prediction comparison in HCP data analysis.

Rank	Node	Node
1	r-lateraloc	r-caudalanterior
2	r-middletemp	r-inferiortemp
3	l-inferiortemp	r-inferiortemp
4	r-caudalanterior	r-precentral

Table 2: Top brain edges selected by our method.

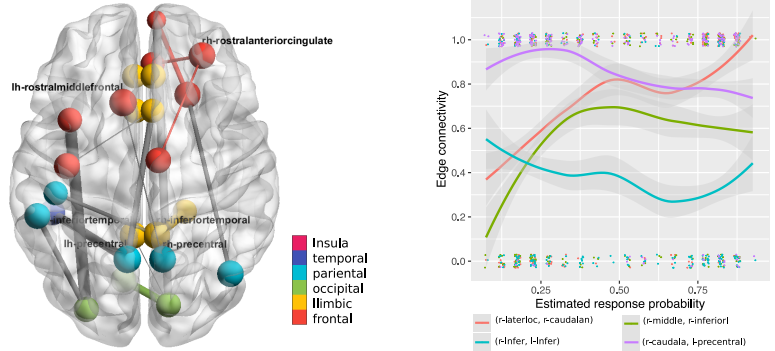


Figure 6: HCP analysis results. (a) Prediction accuracy. (b) Top edges selected by our method. The letter r/l represents the right/left hemisphere. (c) Top edges overlaid on brain template. (d) Edge connectivity versus predicted response probability. The raw connectivity values (0 or 1) are plotted as jitter points. The curve represents the moving average of connectivity value, with one standard error depicted as gray band.

We compare our performance to other methods using the same procedure as in the previous section. Table 1 shows that our method achieves accurate regression, measured by area under receiver operating characteristic (AUC). As common in high-dimensional settings, we observe that models with optimal cross-validation accuracy tend to include many noise variables (Hastie et al., 2015). A useful heuristic is the so-called “one-standard-error rule” (Hastie et al., 2015), in which one selects the most parsimonious model with cross-validation accuracy within one standard error of the best. We apply this rule and report the results as **NonMAR-p**. Remarkable, **NonMAR-p** results in 12% reduction of active nodes but still achieves excellent accuracy (73.6% AUC).

Table 1 lists the top brain edges identified by our method. Edges are ranked by their maximal values in the feature weights $(\hat{\mathbf{B}}_{\pi})_{\pi \in [0,1]}$ via moving average. We find that the top edges mainly

involve connections between frontal and occipital regions in the right hemisphere. This seems consistent with recent evidence of dysfunction in right posterior regions for deficits in visuospatial processing (Wang et al., 2017). We also find nonlinear relationships between edge connections with disease probability. For example, the connection (r-lateraloc, r-caudalanterior) is roughly constant when π is around 0.5, but otherwise has a growing trend. The connection (r-middle, r-inferior) increases when π is low, and then reaches a plateau as π grows. The detected pattern reveals the heterogeneous changes in brain connectivities as the visuospatial processing ability changes.

7 Conclusion

We have developed the learning framework for the relationship between a binary label response and a high-dimensional matrix-valued predictor. The theoretical and numerical results demonstrate the competitive performance of our method. The work unlocks several directions of future research. Application of nonparametric methods to domain problems warrants future research.

References

- Alquier, P. and G. Biau (2013). Sparse single-index model. *Journal of Machine Learning Research* 14 (Jan), 243–280.
- Bartlett, P. L., M. I. Jordan, and J. D. McAuliffe (2006). Convexity, classification, and risk bounds. *Journal of the American Statistical Association* 101(473), 138–156.
- Fan, J., W. Gong, and Z. Zhu (2019). Generalized high-dimensional trace regression via nuclear norm regularization. *Journal of Econometrics*.
- Friedman, J., T. Hastie, and R. Tibshirani (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software* 33(1), 1.
- Ganti, R., N. Rao, L. Balzano, R. Willett, and R. Nowak (2017). On learning high dimensional structured single index models. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pp. 1898–1904. AAAI Press.
- Hastie, T., R. Tibshirani, and M. Wainwright (2015). *Statistical learning with sparsity: the lasso and generalizations*. CRC press.
- Hu, W., W. Shen, H. Zhou, and D. Kong (2020). Matrix linear discriminant analysis. *Technometrics* 62(2), 196–205.
- Relión, J. D. A., D. Kessler, E. Levina, S. F. Taylor, et al. (2019). Network classification with applications to brain connectomics. *The Annals of Applied Statistics* 13(3), 1648–1677.
- Scott, C. and M. Davenport (2007). Regression level set estimation via cost-sensitive classification. *IEEE Transactions on Signal Processing* 55(6), 2752–2757.

- Wang, J., X. Shen, and Y. Liu (2008). Probability estimation for large-margin classifiers. *Biometrika* 95(1), 149–167.
- Wang, L., D. Durante, R. E. Jung, and D. B. Dunson (2017). Bayesian network–response regression. *Bioinformatics* 33(12), 1859–1866.
- Wang, L., Z. Zhang, and D. Dunson (2017). Common and individual structure of multiple networks. *arXiv preprint arXiv:1707.06360*.
- Willelt, R. M. and R. D. Nowak (2007). Minimax optimal level-set estimation. *IEEE Transactions on Image Processing* 16(12), 2965–2979.
- Zhou, H. and L. Li (2014). Regularized matrix regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76(2), 463–483.