UNIVERSITY OF CALIFORNIA,
IRVINE

Statistical Learning for High-dimensional Imaging Data Analysis

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Statistics

by

Wei Hu

Dissertation Committee:
Weining Shen, Chair
Dehan Kong, Cochair
Zhaoxia Yu
Michele Guidani

2019

DEDICATION

To my beloved parents for their unyielding support.

To my mentors and friends for their constant encourage.

TABLE OF CONTENTS

Page

LIST OF FIGURES **v**

LIST OF TABLES **vi**

LIST OF ALGORITHMS **vii**

ACKNOWLEDGMENTS **viii**

CURRICULUM VITAE **ix**

ABSTRACT OF THE DISSERTATION **x**

1 Introduction **1**
- 1.1 Overview of imaging data in applications . . . 1
- 1.2 Overview: common regularization in high-dimensions . . . 3
- 1.3 Existing statistical approaches for imaging data analysis . . . 5
- 1.4 Deep learning on imaging data . . . 8
- 1.5 Contribution . . . 10

2 Matrix Linear Discriminant Analysis **12**
- 2.1 Introduction . . . 12
- 2.2 Method . . . 14
- 2.3 Theory . . . 17
- 2.4 Numerical results . . . 21
  - 2.4.1 Simulation . . . 21
  - 2.4.2 Real data application . . . 24
- 2.5 Discussion . . . 26

3 Nonparametric Matrix Response Regression **27**
- 3.1 Introduction . . . 27
- 3.2 Method . . . 30
  - 3.2.1 Model . . . 30
  - 3.2.2 Bayesian information criterion . . . 33
- 3.3 Theory . . . 35
- 3.4 Simulation . . . 38

3.4.1 Univariate predictor . . . 38
3.4.2 Multivariate predictors . . . 41
3.5 Real data application . . . 43
3.5.1 Application to calcium imaging data study . . . 43
3.5.2 Application to EEG data study . . . 46

4 Latent Representer Values in Image Classification 48
4.1 Introduction . . . 48
4.2 Related Work . . . 50
4.3 Framework . . . 52
4.4 Experiments . . . 58
4.4.1 Dataset Debugging . . . 59
4.4.2 Image exploring . . . 60
4.4.3 Understanding misclassified image . . . 62
4.4.4 Manifold visualization . . . 63
4.5 Discussion . . . 64

Bibliography 66

A Appendix for Chapter 2 75
A.1 Primary lemmas and propositions . . . 75
A.2 Proof of Theorems . . . 80

B Appendix for Chapter 3 86
B.1 Useful lemmas . . . 86
B.2 Curvature and strong convexity . . . 89
B.3 Proof of Theorem 3.1 . . . 91
B.4 Rank consistency: proof of Theorem 3.2 . . . 95

C Appendix for Chapter 4 100
C.1 Supplementary examples of MNIST . . . 100
C.2 Experiments on Fashion-MNIST dataset . . . 101
C.2.1 Image exploring . . . 102
C.2.2 Understanding misclassified images . . . 104
C.2.3 More examples . . . 106

LIST OF FIGURES

Page

2.1 The figures for cross image: (a) original signal; (b) our nuclear regularization estimate; (c) ℓ_1-regularized estimate. 24

3.1 (a)-(c): true signals, (d)-(f): recovered signals 40
3.2 (a) A sequence of frames (b) scatterplot for a fixed voxel of coordinate (200,60) over frames 44
3.3 (a) Fitted value for a fixed voxel of coordinate (200,60) over frames (b) Fitted value for a fixed voxel of coordinate (60,180) over frames 44
3.4 (a) Original 1500th frame (b) Estimated 1500th frame by our method 45
3.5 (a) Estimated 10th frame for alcoholic (b) Estimated 10th frame for control 47

4.1 Bidirectional GAN visualization. 54
4.2 BiGAN training results for permutation-invariant MNIST dataset,including generated samples $G(z)$, real data x from digit 0 to 9, and corresponding reconstructions $G(E(x))$. 59
4.3 Dataset debugging performance of our method and Yeh's method. Our method is able to recover similar amount of flipped training points as Yeh's by inspecting representer value (left) but achieves a far better accuracy (right) after refitting the model with fraction of training data checked as $0.05, 0.10, \cdots, 0.35$. 60
4.4 Comparison of top three excitatory and inhibitory influential training images for a test point(ID 734) (left-most column) using our method (left columns) and Yeh's method (right columns). 61
4.5 Comparison of top three excitatory and inhibitory influential training images for a test point(ID 363) (left-most column) using our method (left columns) and Yeh's method (right columns). 62
4.6 A misclassified test image (left most) and the most influential training point by our method(2nd column) and Yeh's method (4th column) supporting the wrongly predicted label and the most influential training point resisting the true label by our method(3rd column) and Yeh's method (5th column).). 63
4.7 Comparison of t-SNE visualization of recovered training points by our method and Yeh's method for two test point with ID 734(top) and ID 363(bottom). 64

LIST OF TABLES

Page

2.1 Simulation results: misclassification rates (%) and associated standard errors obtained from our method, Lasso LDA, Logistic Nuclear (L-Nuclear), Logistic Lasso (L-Lasso) and penalized matrix discriminant analysis (PMDA) based on 1000 Monte Carlo replications. 22

2.2 EEG data analysis: misclassification rates (%) and associated standard errors. . . . 25

3.1 Simulation results for Setting I: mean of integrated test error and associated standard errors obtained from our method, NW estimator, and Lasso, the average selected rank and true rank are reported for three different shapes B. The results are based on 100 Monte Carlo replications. 40

3.2 Simulation results for Setting II: mean of integrated test error and associated standard errors obtained from our method, NW estimator, and Lasso, the average selected rank and true rank are reported for three different shapes B. The results are based on 100 Monte Carlo replications. 41

3.3 Simulation results for Setting III: mean of integrated test error and associated standard errors obtained from our method, NW estimator, and Lasso, the average selected rank and true rank are reported for three different shapes B. The results are based on 100 Monte Carlo replications. 42

3.4 Simulation results for Setting IV: mean of integrated test error and associated standard errors obtained from our method, NW estimator, and Lasso, the average selected rank and true rank are reported for three different shapes B. The results are based on 100 Monte Carlo replications. 43

3.5 Leave-one-out cross-validation errors (Err)(10^6) and associated standard deviation by three methods for calcium imaging data 45

3.6 Leave-one-out cross-validation errors (SE) by three methods for EEG data 47

List of Algorithms

Page

1 Algorithm to solve the optimization problem (3.4). 33

ACKNOWLEDGMENTS

First and foremost, I would like to express my deepest gratitude to my two advisors Weining Shen and Dehan Kong, who make this thesis possible. Weining and Dehan sparkled my interest in high-dimensional statistical models and its application to neuroimaging. They patiently guided me throughout all aspects including developing statistical methods, building up theoretical framework, programming, scientific writing and presentation skills. These experiences allow me to grow as a researcher. Many thanks to you for your precious help!

I am deeply grateful to my thesis committee members, Zhaoxia Yu, Michele Guindani as well as advance committee members Bin Nan and Hongkai Zhao, for their insightful discussion and feedback, which made me think and dig deeper into my work.

I would like to thank my fellow doctoral students Xu Gao, Tian Chen and Yuxiao Wang for their countless advices and supports. They spend much time sharing with me their research experience and programming techniques, which are greatly helpful when I was a newbie to statistics.

Lastly, I would like to thank all of my friends for their emotional supports. They made me have the courage to face a variety of obstacles and be ready to the next phase of my career.

CURRICULUM VITAE

Wei Hu

EDUCATION

Doctor of Philosophy in Statistics **2019**
University of California, Irvine *Irvine, CA*

Master of Science in Statistics **2017**
University of California, Irvine *Irvine, CA*

Bachelor of Science in Mathematics **2015**
University of Science and Technology of China *Hefei, Anhui*

RESEARCH EXPERIENCE

Graduate Research Assistant **2016–2019**
University of California, Irvine *Irvine, California*

TEACHING EXPERIENCE

Teaching Assistant **2018–2019**
University of California, Irvine *Irvine, CA*

ABSTRACT OF THE DISSERTATION

Statistical Learning for High-dimensional Imaging Data Analysis

By

Wei Hu

Doctor of Philosophy in Statistics

University of California, Irvine, 2019

Weining Shen, Chair

The past two decades have witnessed tremendous advancement in medical imaging techniques. The explosive growth of high-dimensional imaging data brings new challenges to statisticians. Machine learning has opened new horizons in a variety of tasks including image recognition and restoration, personalized medicine, medical image analysis and many others. However, machine learning systems remain mostly black boxes despite widespread adoption. Understanding the statistical properties and the predictions behind black-box models is crucial as it can help to interpret the analysis results. This dissertation dedicates to the development of new statistical learning methods for image data analysis and new insights in understanding block box predictive model behavior.

We start by proposing a novel linear discriminant analysis approach for the classification of high-dimensional matrix-valued data that commonly arises from imaging studies. Motivated by the equivalence of the conventional linear discriminant analysis and the ordinary least squares, we consider an efficient nuclear norm penalized regression that encourages a low-rank structure. Theoretical properties including a non-asymptotic risk bound and a rank consistency result are established. Simulation studies and an application to electroencephalography data show the superior performance of the proposed method over the existing approaches.

Next, we propose a novel nonparametric matrix response regression model to characterize the as-

sociation between 2D image outcomes and predictors such as time and patient information. Our estimation procedure can be formulated as a nuclear norm regularization problem, which can capture the underlying low-rank structures of the dynamic 2D images. We develop an efficient algorithm to solve the optimization problem and introduce a Bayesian information criterion for our model to select the tuning parameters. Asymptotic theories including the risk bound and rank consistency are derived. We finally evaluate the empirical performance of our method using numerical simulations and real data applications from a calcium imaging study and an electroencephalography study.

Finally we propose to trace the predictions of a black-box model back to the training data through a representation theorem calibrated on a continuous, low-dimensional latent space, making the model more transparent. We show that for a given test point and a certain class, the pre-activation prediction value can be decomposed into a sum of representer values, where each representer value corresponds to the importance of the training point on the model prediction. These representer values provide users a deeper understanding of how training points lead the machine learning system to the prediction. We further elaborate our method through theoretical studies, numerical experiments and applications such as debugging models.

Chapter 1

Introduction

1.1 Overview of imaging data in applications

Since the last decade, modern scientific technologies have been producing data with complex structure in the form of matrix or tensor. One main instance is neuroimaging data, which includes calcium imaging, diffusion tensor imaging (DTI), local field potentials (LFPs), functional magnetic resonance imaging (fMRI), electroencephalogram (EEG) and more. fMRI is a class of imaging techniques for measuring regional and temporal metabolic changes in brain. When a region of the brain becomes more active, an increase in blood oxygenation and flow acts to meet the increased energy consumption locally [Glover, 2011]. Imaging the change in blood flow primarily based on blood-oxygen-level dependent (BOLD) contrast produces a three-dimensional image, where each voxel corresponds to a brain location and the value represents changes in magnetic susceptibility. fMRI has promising applications including surgery planning, detecting effects of diseases, monitoring treatment outcomes, as well as some potential ones in translational medicine and clinical practice such as understanding functional brain disorders [Paul et al., 2006].

Calcium imaging is a powerful avenue for observing the spiking activity of large neuronal popu-

lations. Calcium influx into cells occurs whenever a neuron fires and changes in concentration of calcium ions can be detected by the fluorescence of calcium indicator molecules. Therefore the locations of neurons and the times at which they fire can be recorded by a sequence of 2D images taken over the time, with each pixel responding to a continuous value of fluorescent intensity as a result of calcium concentration [Helmchen and Denk, 2005, Petersen et al., 2018]. The primary goal of interest is to locate the major neurons and model their calcium concentration over the time. A detailed discussion of calcium imaging is included in Chapter 3.

Another popular neuroimaging technique, EEG, measures voltage values from an array of electrodes placed on scalps to record brain electrical activity on a variety of temporal frequency, in the sense that between any two electrodes, the difference in voltages can capture current electric flows generated by neurons. Therefore EEG data plays a vital role in numerous clinical applications by connecting brain activity with observed behaviors, such as detecting epileptic seizures [Saab and Gotman, 2005], diagnosing sleep disorders [Platt and Riedel, 2011] and brain computer interface (BCI) [Lotte et al., 2007]. The project described in Chapter 2 is motivated by an EEG data application, where 122 subjects were selected with 77 alcoholic individuals and 45 controls in the aim to study the EEG correlates of genetic predisposition to alcoholism. Each individual's scalp was placed 64 electrodes at 256 Hz (3.9-msec epoch) for 1 second, resulting in a 256×64 matrix as the sampling unit.

Besides imaging data discussed above, a rich source of imaging also includes ultrasound and nuclear medicine imaging. Ultrasonic images, also known as sonograms, are processed by flowing ultrasound pulses using a probe into tissue. The received echos are then transformed into a variety of digital images such as B-mode image, a two-dimensional image of which the brightness reflects echo magnitude. Urtrasound imaging is widely used for diagnosing certain medical conditions especially for pregnancy owing to its safety. Nuclear medicine imaging uses a tiny amount of radioactive substances that are administered to a patient intravenously or orally. A specially designed camera can track and record radiation emitted from the tracers, in a form of two-dimensional

(Scintigraphy) or three-dimensional (SPECT) image.

1.2 Overview: common regularization in high-dimensions

Imaging data bring statistical and computational challenges due to their high and complex structure. This section provides a review of work on various sparse regression models, where the number of predictors p is of the same or greater with the sample size n.

Suppose that we have n observed samples $z_i = (x_i, y_i)$, $i = 1, 2, \cdots, n$ drawing independently from some distribution P_θ and taking values in space $\mathcal{Z}$, where the predictor of interest $x_i = (x_{i1}, x_{i2}, \cdots, x_{in})^T$ is a vector of length p and y_i is the corresponding response. Consider a model called M that maps x_i to y_i by and a cost function $L_n(Z, \theta) : \mathbb{R}^p \times \mathcal{Z}^n \to \mathbb{R}$. which is assumed to be convex and differentiable. The population risk is induced by the expectation of the cost function: $L(\theta) = \mathbb{E}(L_n(Z, \theta))$. Let θ^* be the minimizer of the population risk, i.e.,

$$\theta^* = \arg\min_{\theta \in \mathbb{R}^p} L(\theta).$$

It is common to add a user defined regularization term $R : \mathbb{R}^p \to \mathbb{R}$ to the empirical cost function and solve the following optimization problem:

$$\hat{\theta} = \arg\min_{\theta \in \mathbb{R}^p} L_n(Z, \theta) + \lambda_n R(\theta), \tag{1.1}$$

where $\lambda_n \in \mathbb{R}^+$ is tuning parameter that measures the strength of regularization.

One popular regularization technique is lasso. Lasso was first proposed by Tibshirani [1996]. The

optimization program of lasso for a linear regression model is

$$\hat{\theta} = \arg\min_{\theta \in \mathbb{R}^p} \frac{1}{2n} \sum_{i=1}^{n} (y_i - x_i^T \theta)^2 + \lambda_n \|\theta\|_1, \tag{1.2}$$

where $\|\cdot\|_1$ is l_1 norm, defined as the sum of absolute values of all entries of a vector. The l_1 regularization tends to shrink redundant model parameters into a smaller subset that contributes most to the prediction, owing to which lasso is extensively employed in variable selection and model interpretation. Lasso is then extended to generalized linear model by modifying the objective optimization program by a penalized likelihood loss:

$$\hat{\theta} = \arg\min_{\theta \in \mathbb{R}^p} -\frac{1}{n} \sum_{i=1}^{n} \log p(y_i | x_i) + \lambda_n \|\theta\|_1.$$

Lasso has two main drawbacks. First, under small sample case ($p < n$), lasso selects at most p variables before it saturates. Second, lasso tends to select one variable in a correlated group and ignore others. Zou and Hastie [2005] proposed elastic net regularization to address the limitations by introducing l_2 penalty, i.e.,

$$\hat{\theta} = \arg\min_{\theta \in \mathbb{R}^p} \frac{1}{2n} \sum_{i=1}^{n} (y_i - x_i^T \theta)^2 + \lambda_1 \|\theta\|_1 + \lambda_2 \|\theta\|_2.$$

Many other extensions of lasso includes SCAD, fussed lasso, grouped lasso, adaptive lasso and so on fitting on a variety of structure of model parameters [Fan and Li, 2001, Tibshirani et al., 2005, Yuan and Lin, 2006, Zou and Hastie, 2005, Zou, 2006]. Besides regression, as to binary classification purpose, the equivalence between ordinary least squares (OLS) and linear discriminant analysis (LDA) motivates some work on developing sparse LDA using l_1-regularization, which boils down to the scope of lasso.

The aforementioned methods deal with vector covariates. However, in imaging analysis, the data can often be represented as a 2d matrix or a 3D tensor. Conventional approaches often stack the

image input into a ultra-long vector and apply the Lasso method. This has two main drawbacks. On the one hand, simple vectorization destroys structural information within image covariates. On the other hand, the l_1 regularization often relies on the sparsity assumption of the underlying parameters, which may not hold for imaging data.

1.3 Existing statistical approaches for imaging data analysis

In this section, we review some related works on (generalized) linear models for matrix-valued data. In real data application, it is often the case that low rankness is a more reasonable assumption than sparseness assumption for the true signal. Analogous to lasso regularization that encourages sparsity for the parameter space, nuclear (spectral) regularization penalizes on the sum of singular values of a matrix, therefore forces the estimator to hold a low-rank structure. To start with, we restate a singular value thresholding algorithm in [Cai et al., 2010] that is useful for solving optimization program involving nuclear norm regularization.

THEOREM 1.1. *Given a matrix $\boldsymbol{X} \in \mathbb{R}^{p\times q}$ with rank r having singular value decomposition:*

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*, \quad \boldsymbol{\Sigma} = diag(\{\sigma_{\mathbf{i}}\}_{\mathbf{i=1}}^{\mathbf{r}}),$$

where $\boldsymbol{U} \in \mathbb{R}^{p\times r}$ and $\boldsymbol{V} \in \mathbb{R}^{r\times q}$ are matrices with orthonormal columns. Denote $\|\cdot\|_*$ to be the nulear norm. The optimal solution to

$$\min_{\boldsymbol{X}} \frac{1}{2}\|\boldsymbol{X}-\boldsymbol{Y}\|_F^2 + \lambda\|\boldsymbol{X}\|_*$$

is $\boldsymbol{U}D_\lambda(\boldsymbol{\Sigma})\boldsymbol{V}^*$, where $D_\lambda(\boldsymbol{\Sigma}) = diag(\{(\sigma_{\mathbf{i}}-\lambda)^+\})$.

The singular value thresholding algorithm to some extent indicates that adding nuclear norm regularization is equivalent to shrinking the singular values of estimator by the regularization with

strength related to λ.

Zhou and Li [2014a] introduced a class of regularized regression methods with matrix covariates hinging upon spectral regularization. Consider a linear regression situation and denote the matrix of parameters by $\mathbf{B} \in \mathbb{R}^{p\times q}$ and the matrix covariate by $\mathbf{X} \in \mathbb{R}^{p\times q}$. The objective optimization problem is formulated as

$$\min \frac{1}{2}\sum_{i=1}^{n}(y_i - \gamma^T z_i - < \mathbf{X}_i, \mathbf{B} >)^2 + \lambda_n \|\mathbf{B}\|_*, \tag{1.3}$$

where $< \cdot, \cdot >$ denotes the inner product of two matrices. z_i is a vector of covariates corresponding to ith observation. Suppose $\mathbf{B}$ has singular value decompostion $\mathbf{U\Sigma V}^*$ where $\mathbf{\Sigma}$ is a diagonal matrix with positive diagonal values $\sigma_1(\mathbf{B}) \geq \sigma_2(\mathbf{B}) \geq \cdots \geq \sigma_r(\mathbf{B}) > 0$. Then $\|\mathbf{B}\|_* = \sum_{j=1}^{r} \sigma_j$, i.e., the problem becomes an l_1-norm regularization on singular values pushing estimated $\mathbf{B}$ into a low-rank structure. The spectral regularization formulation is extended to GLM framework straightforwardly by replacing the squared loss with negative log-likelihood function.

Different from lasso, spectral regularization is convex. Therefore any local minima of optimization program (1.3) is global minimum. Zhou and Li [2014a] proposed a numerical solution of (1.3) using the Nesterov optimal gradient algorithm, which applies singular value thresholding formula [Cai et al., 2010] in each iteration and combines two algorithmic iterates to update the estimate.

Still consider a scalar response and a matrix-valued predictor but the matrix of parameters is piecewise smooth with unknown edges and jumps rather than a low-rank structure. The assumption of piecewise smoothness is widely adopted in imaging studies such as associating brain regions with Alzheimers disease. Wang et al. [2017] proposed a generalized scalar-on-image regression models with penalization via total variation (GSIRM-TV), which was demonstrated to be potent in preserving the boundaries of images. Under the simplest case of GSIRM-TV, a linear scalar-on-image

regression model, the loss function is given by

$$\sum_{i=1}^{n}(Y_i - < \mathbf{X}_i, \mathbf{B} >)^2 + \lambda_n \|\mathbf{B}\|_{\text{TV}},$$

where $\|\mathbf{B}\|_{\text{TV}}$ is the total variation of $\mathbf{B}$, i.e.,

$$\|\mathbf{B}\|_{\text{TV}} = \sup\{\int_{\omega} \mathbf{B}(u,v)\text{div}f(u,v)dudv : f \in \mathcal{C}_c^{\infty}(\Omega;\mathbb{R}^2), |f|_{\infty} \leq 1\}. \tag{1.4}$$

The div in (1.4) is the divergence operator defined on a vector field. The total variation regularization encourages $\mathbf{B}$ to be piesewise smooth and meanwhile is capable of preserving sharp boundaries and jumps. The minimizer of (1.4) can be solved iteratively where each iterate is updated by the augmented Lagrangian method.

Besides the situation where the response is a scalar or a vector, the response can also take a matrix form that characterizes structural connectivity pattern in imaging genetics and the corresponding covariates are of a vector including age, gender and many others. Kong et al. [2019] investigated a low-rank linear regression model with high-dimensional matrix response and high dimensional scalar covariates by considering

$$\mathbf{Y}_i = \sum_{l=1}^{s} x_{il} * \mathbf{B}_l + \mathbf{E}_i,$$

where Y_i is $p \times q$ matrix of response and x_{il} is a vector of s covariates. To recover the low-rank structure of coefficient matrices, the loss function is

$$\frac{1}{2n}\sum_{i=1}^{n} \|\mathbf{Y}_i - \sum_{l\in\mathcal{M}} x_{il} * \mathbf{B}_l\|_F + \lambda \sum_{i\in\mathcal{M}} \|\mathbf{B}_l\|_*, \tag{1.5}$$

where $\mathcal{M}$ is the set of indices of covariates after pre-screening and $\|\cdot\|_F$ is the Frobenius norm. Solving the minimizer of the objective function (1.5) is based on Nesterov gradient method and

singular value thresholding formula [Cai et al., 2010].

1.4 Deep learning on imaging data

Deep learning methods have gained an increasing attention in medical imaging owing to their capacity in learning underlying complex data distributions. Potential applications include data augmentation for tumor detection [Han et al., 2019], data completion for improving disease diagnosis [Li et al., 2014], personalized treatment suggestions [Nezhad et al., 2016, Krittanawong et al., 2017], medical image synthesis [Nie et al., 2017] and many others. Among a variety of deep learning methods, Variational Auto-Encoders (VAEs) [Kingma and Welling, 2013] and Generative Adversarial Networks (GANs) [Goodfellow et al., 2014] are two breakthroughs in deep generative models in the sense of approximating data distribution defined on some high-dimensional space and generating complicated synthetic images.

In VAE, the goal is to maximize $p(\mathbf{x}) = \int p_\theta(\mathbf{x}|\mathbf{z})p(\mathbf{z})d\mathbf{z}$, the distribution of datapoints $\mathbf{x}$ generated from $\mathbf{z}$ in some low-dimensional space $\mathcal{Z}$. However, approximating $\int p_\theta(\mathbf{x}|\mathbf{z})p(\mathbf{z})d\mathbf{z}$ by $\frac{1}{n}\sum_{i=1}^{n} p_\theta(\mathbf{x}|\mathbf{z}^i)$ is impractical since $p_\theta(\mathbf{x}|\mathbf{z}^i)$ is almost 0 for most values of $\mathbf{z}$. Therefore it is of interest to learn a distribution $q_\phi(\mathbf{z}|\mathbf{x})$ that is more likely to generate $\mathbf{x}$ [Doersch, 2016]. To connect $q_\phi(\mathbf{z}|\mathbf{x})$ with $p(\mathbf{x})$, the following relationship holds:

$$\log p(\mathbf{x}) - D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})\|p(\mathbf{z}|\mathbf{x})) = E_{q_\phi}(\log p_\theta(\mathbf{x}|\mathbf{z})) - D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})\|p_\theta(\mathbf{z})). \tag{1.6}$$

Therefore the objective is to maximize the right side of 1.6 and at the same time to minimize the KL-divergence between $q_\phi(\mathbf{z})$ and $p(\mathbf{z}|\mathbf{x})$. Since the true posterior $p(\mathbf{z}|\mathbf{x})$ is always unknown, when the model $q_\theta(\mathbf{z}|\mathbf{x})$ has high capacity, $D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})\|p(\mathbf{z}|\mathbf{x}))$ is fairly small and then can be ignored. As a result, the right side of (1.6) is the ultimate object to be optimized, called Evidence Lower Bound (ELBO). The usual candidate of $q_\phi(\mathbf{z}|\mathbf{x})$ is $N(\mathbf{z}|\mu(\mathbf{x}), \Sigma(\mathbf{x}))$ where $\mu(\mathbf{x})$ and $\Sigma(\mathbf{x})$ are some

neural network outputs. Here $p_\theta(\mathbf{x}|\mathbf{z})$ is a function of $\mathbf{z}$ denoted by $f(\theta, \mathbf{z})$, and $q_\phi(\mathbf{z}|\mathbf{x})$ essentially acts as an encoder that maps $\mathbf{x}$ to its latent representation $\mathbf{z}$ and $p_\theta(\mathbf{x}|\mathbf{z})$ is a decoder that decodes $\mathbf{z}$ to reconstruct $\mathbf{x}$. Optimizing process is then possible through stochastic gradient descent via backpropagation with reparameterization trick of sampling.

Compared to VAE equipped with explicit density function, GANs is a class of "likelihood free" generative models, i.e., no explicit density function is required, which allows more flexibility. The GAN framework consists of two neural networks: generator and discriminator. The generator $G : \mathcal{Z} \to \mathcal{X}$ is a deconvolutional neural network that generates synthetic images from noise $\mathbf{z}$ sampled from some low-dimensional latent space. The discriminator $D : \mathcal{X} \to [0, 1]$ is trained to distinguish between synthetic and real images. Two networks compete with each other in the way that the generator is trained to generate images that are close to real ones to fool the discriminator. Denote the prior of noise $z \in \mathcal{Z}$ to be $p_\mathbf{z}(\mathbf{z})$ and the distribution of data $x \in \mathcal{X}$ to be $p_\mathbf{x}(\mathbf{x})$. Both generator and discriminator are trained simultaneously via the adversarial objective $\min_G \max_D V(D, G)$, where

$$V(D, G) = E_{\mathbf{x}\sim p_\mathbf{x}(\mathbf{x})}(\log(D(\mathbf{x})) + E_{\mathbf{z}\sim p_\mathbf{z}(\mathbf{z})}(\log(1 - D(G(\mathbf{z})))), \tag{1.7}$$

and $D(\mathbf{x})$ is the probability that $\mathbf{x}$ is sampled from $p_\mathbf{x}(\mathbf{x})$ rather than generated by G. The minimax objective is intuitive considering that the generator tries to fool the discriminator but the discriminator aims to distinguish real and fake images as accurately as possible. Goodfellow et al. [2014] demonstrated that when G is fixed, the optimal discriminator is

$$D^*_G(\mathbf{x}) = \frac{p_{data}(\mathbf{x})}{p_{data}(\mathbf{x}) + p_g(\mathbf{x})}.$$

Given the optimal discriminator $D^*_G(\mathbf{x})$, it is trivial to show that the minimax objective is

$$V(G, D^*_G(\mathbf{x})) = 2D_{JSD}[p_{data}, p_G] - \log 4,$$

where D_{JSD} is Jenson-Shannon Divergence. Since D_{JSD} is always non-negative, the optimal generator is reached when $p_{data} = p_G$. GAN has shown superior performance in a variaty of tasks, however, training GAN remains challenging in practice due to unstable optimization and mode collapse. There have been numerous works focusing on improving GAN's training process including Salimans et al. [2016], Heusel et al. [2017], Salimans et al. [2018] and many others.

1.5 Contribution

Our contribution of this dissertation is multifold.

First, in Chapter 2, we propose a novel matrix linear discriminant analysis (LDA) approach for the classification of high-dimensional matrix-valued data. The binary classification was formulated as a penalized least-squares problem with nuclear norm regularization, which efficiently exploits the low-rank structure of the two-dimensional discriminant direction matrix. We also derive the risk bound of the estimator, which is explicit in terms of the rank of the image, image size, sample size, and the eigenvalues of the covariance matrix for the image covariates. We show that to achieve estimation consistency given $p \times q$ image, one sufficient condition is that $\max(p, q) = o(n/\log^3 n)$. Under stronger conditions, the estimated rank of the coefficient matrix is proved to be consistent as well. Finally, we prove that our method enjoys classification error consistency.

Chapter 3 provides a novel regression approach to model the association between 2D image response and predictors such as time, patient demographics, and other disease predictors. In contrast to the dominating use of linear models in the literature, we adopt a flexible nonparametric regression model to capture the commonly-seen nonlinear relationship in the data. For choosing optimal tuning parameter, an analytic form of BIC was derived, which is not straightforward in our nonparametric matrix response model. Finally, we develop the asymptotic theories including the risk bound and rank consistency for the proposed nonparametric estimator, which directly connects to

the existing work on nonparametric statistics theory.

Chapter 4 focuses on interpreting predictions of black-box models arising in complex machine learning systems. We present a representer theorem calibrated on latent space, which decomposes the pre-activation value into a sum of representer values of training points. By the sign and significance of represent value, we are able to identify a training point to be an excitory or inhibitory to a specific prediction, aiding a richer understanding towards the model's behavior. We also demonstrate that the representer values measured on latent space characterize the "true" influence of each training point on the prediction in the sense of approximating the true ranking of influences of training data. The superior performance of our method is substantiated by adequate heuristic experiments on large-scale datasets.

Chapter 2

Matrix Linear Discriminant Analysis

2.1 Introduction

Modern technologies have generated a large number of datasets that possess a matrix structure for classification purpose. For example, in neuropsychiatric disease studies, it is often of interest to evaluate the prediction accuracy of prognostic biomarkers by relating two-dimensional imaging predictors, e.g., electroencephalography (EEG) and magnetoencephalography, to clinical outcomes such as diagnostic status [Mu and Gage, 2011]. In this paper, we focus on extending one of the most commonly used classification methods, Fisher linear discriminant analysis (LDA) to matrix-valued predictors. Progress has been made in recent years on developing sparse LDA using ℓ_1-regularization [Tibshirani, 1996], including Shao et al. [2011], Fan et al. [2012], Mai et al. [2012]. However, all these methods only deal with vector-valued covariates; and it remains challenging to accommodate the matrix structure. Naively transforming the matrix data into a high-dimensional vector will result in unsatisfactory results for several reasons. First, vectorization destroys the structural information within the matrix such as shapes and spatial correlations. Second, turning a $p \times q$ matrix into a $pq \times 1$ vector generates unmanageably high dimensionality.

E.g., estimating the population precision matrix for LDA can be troublesome if $pq \gg n$. Third, ℓ_1-regularization does not necessarily work well because the underlying two-dimensional signals are usually approximately low-rank rather than ℓ_0-sparse.

Recently, there are some development of regression methods for matrix data. Zhou and Li [2014a] proposed a class of regularized matrix regression methods based on spectral regularization. Wang and Zhu [2017] developed a generalized scalar-on-image regression model via total variation. Chen et al. [2013b] invented an adaptive nuclear norm penalization approach for low-rank matrix approximation.

In this paper, we propose a new matrix LDA approach by building on the equivalence between the classical LDA and the ordinary least squares. We formulate the binary classification as a nuclear norm penalized least squares problem, which efficiently exploits the low rank structure of the two-dimensional discriminant direction matrix. The involved optimization is amenable to the accelerated proximal gradient method. Although our problem is formulated as a penalized regression problem, a fundamental difference is that the covariates $\mathbf{X}_i$ and the residuals ϵ_i are no longer independent in our case. This requires extra effort for developing the risk bound and rank consistency result. The risk bound is explicit in terms of the rank of the image, image size, sample size, and the eigenvalues of the covariance matrix for the image covariates. This result also implies estimation consistency provided the $p \times q$ image satisfies $\max(p, q) = o(n/\log^3 n)$. Under stronger conditions, we show that the rank of the coefficient matrix can be consistently estimated as well. The proof is based on exploiting the spectral norm of random matrices with mixture-of-Gaussian components and extending the results in Bach [2008a] to allow diverging matrix dimensions. Finally, we prove that our method enjoys classification error consistency.

It is worth noting that the 2D image classification problem has been studied by Zhong and Suslick [2015], where they proposed a penalized matrix discriminant analysis method (PMDA) that projects the matrix coefficient into row space and column space separately. Those two projections are then estimated iteratively and integrated together for classification. Compared with PMDA,

we make the following contributions. First, the rank of the PMDA is set as one because of the separability assumption, while we allow the rank of the direction matrix to take general positive integer values and the rank can then be selected by a data driven procedure. Our rank assumption is more flexible in practice and hence often leads to a lower mis-classification error in the numerical studies. Second, our method adopts a direct estimation approach by solving a nuclear norm penalized regression problem, which is computationally much faster compared with PMDA, where the estimation involves an iterative procedure for calculating the inverse of covariance matrices during each iteration. Third, our method can handle the high-dimensional data when image dimensions p and q are much larger than the sample size, which is the case for many applications; while PMDA cannot handle the case when $p + q > n$. Finally, we have provided theoretical guarantee for our estimator when p and q diverge with n. In particular, we have developed an non-asymptotic error bound for the estimated LDA direction, as well as results on rank consistency and classification error consistency. These results are stronger compared with the root-n consistency of the LDA direction in Zhong and Suslick [2015], where both p and q are assumed to be fixed.

2.2 Method

We first define some useful notations. Let $\text{vec}(\cdot)$ be a vectorization operator, which stacks the entries of a matrix into a column vector. The inner product between two matrices of same size is defined as $\langle \mathbf{M}, \mathbf{N} \rangle = \text{tr}(\mathbf{M}^{\mathrm{T}}\mathbf{N}) = \langle \text{vec}(\mathbf{M}), \text{vec}(\mathbf{N}) \rangle$.

Consider a binary classification problem, where $\mathbf{X}$ is a two-dimensional image covariate with dimension $p \times q$ and $G = 1, 2$ denotes the class labels. The LDA assumes that $\text{vec}(\mathbf{X}) \mid G = g \sim N(\mu_g, \mathbf{\Sigma})$, $\text{pr}(G = 1) = \pi_1$, and $\text{pr}(G = 2) = \pi_2$. Suppose we have n subjects with n_1 subjects belonging to class 1 and $n_2 = n - n_1$ subjects to class 2. It is well known that LDA is connected to the linear regression with the class labels as responses [Duda et al., 2012, Mika, 2002]. When

$pq < n$, the classical LDA is equivalent to solving

$$(\hat{\beta}_0^{\text{ols}}, \hat{\mathbf{B}}^{\text{ols}}) = \arg\min_{\beta_0, B} \sum_{i=1}^{n} \Big(y_i - \beta_0 - \langle \mathbf{X_i}, \mathbf{B} \rangle \Big)^2, \tag{2.1}$$

where $\mathbf{X}_i$ is the image covariate from subject i, $\mathbf{B}$ is the coefficient matrix for the image covariate and it represents the direction of the linear discriminant classifier, β_0 is the intercept, and the response $y_i = -n/n_1$ if subject i is in class 1, and $y_i = n/n_2$ if subject i is in class 2. Although this connection gives the exact LDA direction when $pq < n$, it has two potential drawbacks. First, when $pq > n$, the equivalence between Fisher LDA and (2.1) is lost because of the non-uniqueness of solution. Second, the formulation (2.1) does not incorporate the 2D image structure when estimating the direction because $\langle \mathbf{X}_i, \mathbf{B} \rangle = \langle \text{vec}(\mathbf{X}_i), \text{vec}(\mathbf{B}) \rangle$. These motivate us to consider a penalized version of (2.1) as follows

$$(\hat{\beta}_0, \hat{\mathbf{B}}) = \arg\min_{\beta_0, \mathbf{B}} \frac{1}{2n} \sum_{i=1}^{n} \Big(y_i - \beta_0 - \langle \mathbf{X}_i, \mathbf{B} \rangle \Big)^2 + \omega_n \|\mathbf{B}\|_*, \tag{2.2}$$

where the nuclear norm $\|\mathbf{B}\|_* = \sum_j \sigma_j(\mathbf{B})$ and $\sigma_j(\mathbf{B})$s are the singular values of the matrix $\mathbf{B}$. The nuclear norm $\|\mathbf{B}\|_*$ plays an important role because it imposes a low rank structure in the estimated direction $\hat{\mathbf{B}}$. An alternative choice is to add a Lasso type penalty, i.e. $\omega_n \|\mathbf{B}\|_{1,1} = \omega_n \sum_{j=1}^{p} \sum_{k=1}^{q} |b_{jk}|$, where b_{jk} is the jk-th element of $\mathbf{B}$. However, the Lasso type penalty can only identify at most n nonzero components, and for most cases in imaging studies, the signal is usually not that sparse. More importantly, the Lasso type of penalty ignores the matrix structure because it is equivalent to vectorizing the array and applying sparse LDA. Once $\hat{\mathbf{B}}$ from (2.2) is obtained, a naive classification rule will assign the i-th subject to class 2 if $\langle \mathbf{X}_i, \hat{\mathbf{B}} \rangle + \hat{\beta}_0 > 0$. However, it can be shown that the intercept $\hat{\beta}_0$ obtained from (2.2) is not optimal. Instead, we use the optimal intercept $\tilde{\beta}_0$ that minimizes the training error after obtaining $\hat{\mathbf{B}}$. Mai et al. [2012] showed that the intercept of LDA actually has a closed form. Their derivations can be easily

applied to our case. In particular, if $(\hat{\mu}_2 - \hat{\mu}_1)^{\mathrm{T}}\mathrm{vec}(\hat{\mathbf{B}}) > 0$, then

$$\tilde{\beta}_0 = -(\hat{\mu}_1 + \hat{\mu}_2)^{\mathrm{T}}\mathrm{vec}(\hat{\mathbf{B}})/2 + \mathrm{vec}(\hat{\mathbf{B}})^{\mathrm{T}}\hat{\mathbf{\Sigma}}\mathrm{vec}(\hat{\mathbf{B}})\{(\hat{\mu}_2 - \hat{\mu}_1)^{\mathrm{T}}\mathrm{vec}(\hat{\mathbf{B}})\}^{-1}\log(n_2/n_1), \quad (2.3)$$

where $\hat{\mu}_g$ is the sample mean for subjects in class g and $\hat{\Sigma}$ is the estimated covariance matrix. If $(\hat{\mu}_2 - \hat{\mu}_1)^{\mathrm{T}}\mathrm{vec}(\hat{\mathbf{B}}) < 0$, we can plug $-\hat{\mathbf{B}}$ into (2.3) to obtain the optimal intercept $\tilde{\beta}_0$. The optimal classification rule is to assign the i-th subject to class 2 if $\langle \mathbf{X}_i, \hat{\mathbf{B}} \rangle + \tilde{\beta}_0 > 0$.

For any fixed ω_n, the optimization problem in (2.2) can be solved using the accelerated proximal gradient method [Nesterov, 1983, Beck and Teboulle, 2009]. Zhou and Li [2014a] studied the algorithm for the nuclear norm regularized matrix regression. As we know, nuclear norm is not differentiable. Fortunately, its subderivative $\partial\|.\|_*$ exists. Therefore (2.2) has local minima $(\hat{\beta}_0, \hat{\mathbf{B}})$ if and only if $0 \in -\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\epsilon_i + \omega_n\partial\|\hat{\mathbf{B}}\|_*$. Thanks to the convexity of nuclear norm, the local minima is global as well. Based on these facts, singular value thresholding method for nuclear norm regularization was deployed for building blocks of the Nesterov's method. Compared with classical gradient decent method with convergence of $O(t^{-1})$, where t denotes the number of iteration, Nesterov's accelerated gradient decent method achieves convergence rate of $O(t^{-2})$. It differs from traditional algorithms by utilizing the estimators from previous two iterations to generate the next estimator. For computational algorithm, we use the `matrix_sparsereg` function in the Matlab TensorReg Toolbox (`https://hua-zhou.github.io/TensorReg/`) for solving nuclear norm penalized matrix regression. It implements an optimal Nesterov acceleration of the proximal gradient algorithm. Actually one contribution of our paper is to link matrix LDA to regularized matrix regression so that the computational machinery developed for the latter can be applied to matrix LDA problems. For tuning of the ω_n, we adopt the BIC derived by Zhou and Li [2014a] under the nuclear norm regularized matrix regression framework.

2.3 Theory

In this section we discuss the theoretical properties of the proposed regularization estimator. Denote the residuals $\epsilon_i = y_i - \beta_0 - \langle \mathbf{X}_i, \mathbf{B} \rangle$ and the true coefficient matrix by $\mathbf{B}_0$. By the equivalence between LDA direction and least squares, we know $\mathrm{vec}(\mathbf{B}_0)$ can be written as $c\mathbf{\Sigma}^{-1}(\mu_2 - \mu_1)$ for some positive constant c. Consider the singular value decomposition $\mathbf{B}_0 = \mathbf{U}_0 \mathrm{Diag}(S_0)\mathbf{V}_0^{\mathrm{T}}$ with $\mathbf{U}_0 \in \mathrm{R}^{p\times r}$ and $\mathbf{V}_0 \in \mathrm{R}^{q\times r}$. Let $\mathbf{U}_{0\perp} \in \mathrm{R}^{p\times(p-r)}$ and $\mathbf{V}_{0\perp} \in \mathrm{R}^{q\times(q-r)}$ be (arbitrary) orthogonal complements of $\mathbf{U}_0$ and $\mathbf{V}_0$, respectively. We make the following assumptions.

(A1) We assume that the second-order moment of the covariate $\mathbf{X}$, $\mathrm{E}(\mathrm{vec}(\mathbf{X})\mathrm{vec}(\mathbf{X})^{\mathrm{T}})$, denoted by $\mathbf{\Sigma}_{xx}$, satisfies $\lambda_l \leq \lambda_{\min}(\mathbf{\Sigma}_{xx}) \leq \lambda_{\max}(\mathbf{\Sigma}_{xx}) \leq \lambda_u$, where $\lambda_{\min}(\mathbf{\Sigma}_{xx})$ and $\lambda_{\max}(\mathbf{\Sigma}_{xx})$ are the smallest and largest eigenvalues of $\mathbf{\Sigma}_{xx}$, respectively, and λ_l, λ_u are some positive constants.

(A2) Let $r = \mathrm{rank}(\mathbf{B}_0)$ be the unknown rank of the true coefficient matrix $\mathbf{B}_0$. Define $\mathbf{\Lambda} \in \mathrm{R}^{(p-r)\times(q-r)}$ as

$$\mathrm{vec}(\mathbf{\Lambda}) = \{(\mathbf{V}_{0\perp} \otimes \mathbf{U}_{0\perp})^{\mathrm{T}}\mathbf{\Sigma}^{-1}(\mathbf{V}_{0\perp} \otimes \mathbf{U}_{0\perp})\}^{-1}\{(\mathbf{V}_{0\perp} \otimes \mathbf{U}_{0\perp})^{\mathrm{T}}\mathbf{\Sigma}^{-1}(\mathbf{V}_0 \otimes \mathbf{U}_0)\mathrm{vec}(\mathbf{I})\}.$$

We assume its spectral norm $\|\mathbf{\Lambda}\|_2 < 1$.

(A3) Assume the quantities ω_n, $\{\min(p,q)\}^{1/2}n^{-1/2}\omega_n^{-1}$, $\min(p,q)n^{-1/2}$, $\omega_n p^{1/2}q^{1/2}\min(p,q)$ tend to 0 as $n \to \infty$.

(A4) There exists a positive constant C_μ such that $\|\mu_2 - \mu_1\|_2 \leq C_\mu(\sqrt{p} + \sqrt{q})$.

Condition (A1) requires bounded eigenvalues for the covariance matrix of the vectored covariate, which is standard in the literature. Condition (A2) is similar with the strict consistency condition in Bach [2008a]. It is needed to establish rank consistency. This condition extends the classical strong irrepresentable condition in Zhao and Yu [2006], which is commonly used for proving

model selection consistency of Lasso. The major difference between our Assumption (A2) and the similar assumption in Bach [2008a] is that the number of parameters is fixed in Bach [2008a] while in our case the number is diverging with n. Therefore we will need to assume that the regularization parameter ω_n decays slower than the one in Bach [2008a]. Condition (A3) puts more requirement on the order of p, q, and w_n in order to obtain consistent rank estimation in addition to consistent coefficient estimation. This is expected since rank estimation consistency is usually not implied by parameter estimation consistency. Condition (A4) can be viewed as a sparsity assumption on B_0. Recall the solution (the slope) to classical LDA problem with vector covariates depends on the term $\mu_2 - \mu_1$. This assumption essentially implies that there are at most $O(\max(p, q))$ number of $O(1)$ elements in the true coefficient matrix $\mathbf{B}_0$ given the rank of $\mathbf{B}_0$ is fixed.

Next, we briefly review two important concepts, namely decomposable regularizer and strong convex loss function, proposed by Negahban et al. [2012] and highlight their connection to the risk bound property for our estimator.

DEFINITION 2.1. *A regularizer $R(\cdot)$ is decomposable with respect to a given pair of subspaces (M, N) where $M \subseteq N^{\perp}$ if*

$$R(u + v) = R(u) + R(v) \quad \textit{for all } u \in M, v \in N.$$

In our setting, $R(\cdot)$ is the nuclear norm. Considering a matrix $\mathbf{B} \in \mathcal{R}^{p \times q}$ to be estimated, we observe that nuclear norm is decomposable given a pair of subspaces:

$$M(\mathbf{U}, \mathbf{V}) := \{\mathbf{B} \in \mathcal{R}^{p \times q} \mid \text{row}(\mathbf{B}) \subseteq \mathbf{V},\ \text{col}(\mathbf{B}) \subseteq \mathbf{U}\},$$

$$N(\mathbf{U}, \mathbf{V}) := \{\mathbf{B} \in \mathcal{R}^{p \times q} \mid \text{row}(\mathbf{B}) \subseteq \mathbf{V}^{\perp},\ \text{col}(\mathbf{B}) \subseteq \mathbf{U}^{\perp}\},$$

where $\mathbf{U}, \mathbf{V}$ represent B's left and right singular vectors. For any pair of matrices $\mathbf{B}_1 \in M$ and $\mathbf{B}_2 \in N$, the inner product of $\mathbf{B}_1, \mathbf{B}_2$ is 0 due to their mutually orthogonal rows and columns.

Hence we conclude $R(\mathbf{B}_1 + \mathbf{B}_2) = R(\mathbf{B}_1) + R(\mathbf{B}_2)$. Since we assume the true parameter has a low rank structure, we expect the regularized estimator to have a large value of projection on $M(\mathbf{U}, \mathbf{V})$ and a relatively small valued projection on $N(\mathbf{U}, \mathbf{V})$.

When the loss function $L(\hat{\beta}_0, \hat{\mathbf{B}}_{\omega_n})$ defined as $\frac{1}{2n}\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \langle \mathbf{X}_i, \hat{\mathbf{B}}_{\omega_n}\rangle\right)^2$ is close to $L(\beta_0, \mathbf{B}_0)$, it is insufficient to claim $\hat{\mathbf{B}}_{\omega_n} - \mathbf{B}_0$ is small if the loss function L is relatively flat. This is why the strong convexity condition is required.

DEFINITION 2.2. *For a given loss function L and norm $\|.\|$, we say L is strong convex with curvature k_L and tolerance function τ_L if*

$$\delta L(\mathbf{\Delta}, \mathbf{B}_0) \geq k_L \|\mathbf{\Delta}\|^2 - \tau_L^2(\mathbf{B}_0), \quad \textit{for any } \delta \in \mathcal{C}(M, N; \mathbf{B}_0),$$

where $\mathcal{C}(M, N; \mathbf{B}_0) := \{\mathbf{\Delta} \in \mathcal{R}^{p\times q} \mid R(\mathbf{\Delta}_N) \leq 3R(\mathbf{\Delta}_{N^\perp}) + 4R(\mathbf{B}_{0N})\}$.

Now we are ready to state the main result on the risk bound for our estimate. The proof is provided in the Appendix B.

THEOREM 2.1. *Suppose that (A1) and (A4) hold. Let $\hat{\mathbf{B}}$ be the solution to* (2.2)*. If*

$$\omega_n \geq \frac{12(\log n)^{3/2}(C_\mu + \lambda_u^{1/2})(\sqrt{p} + \sqrt{q} + \sqrt{\log n})}{\sqrt{n}},$$

then with probability of at least $1 - Cn^{-1}$ for some constant $C > 0$,

$$\|\hat{\mathbf{B}} - \mathbf{B}_0\|_F^2 + |\hat{\beta} - \beta_0^*|^2 \leq 9\frac{\omega_n^2}{\lambda_l} r,$$

where $\beta_0^* = \beta_0 - \pi_2^{-1}\{c - 1 + (\pi_2 - \pi_2^2)(\mathbf{D}^{\mathrm{T}}\mathbf{\Sigma}^{-1}\mathbf{D})\}$ and c is some positive constant.

Theorem 2.1 gives a non-asymptotic risk bound for the proposed estimators. In other words, the results hold for any positive ω_n satisfying the conditions there. However, in order to ensure the consistency of the proposed estimators, we will need the risk bound to go to 0, which requires

$\omega_n \to 0$ and $\max(p,q) = o\left(n/(r\log^3 n)\right)$. If the rank of $\mathbf{B}_0$ is fixed, then both p and q can diverge with n at the order of $o(n/\log^3 n)$ and their product $pq > n$. This result is compatible with Theorem 1 in Raskutti and Yuan [2015]. Note that the estimated intercept $\hat{\beta}$ converges to β_0^*, which deviates from the truth β_0. This is expected because the solution to OLS is only equivalent with LDA's solution in terms of the slope $\mathbf{B}$, not on β_0. More precisely, for OLS, by taking the derivative of squared loss function with respect to β_0 and set it to 0, we essentially require $E(\epsilon) = 0$. However, this does not hold in our case. Instead we need to shift the residual ϵ by d to balance off the bias in the cross-product term $\mathrm{E}(\epsilon\mathbf{X})$. The proof of the theorem uses Gaussian comparison inequality which allows us to deal with $\mathrm{vec}(\mathbf{X})$ following a general Gaussian distribution instead of standard Gaussian distribution given that the largest singular value of $\mathbf{\Sigma}_{xx}$ is bounded. Based on this connection, we further utilize concentration property of spectral norm of Gaussian random matrices.

Next we show that $\hat{\mathbf{B}}$ is rank-consistent under stronger conditions.

THEOREM 2.2. *Suppose that (A1)–(A4) hold. Then the estimate* $\hat{\mathbf{B}}$ *is rank-consistent, that is,* $P(\textit{rank}(\hat{\mathbf{B}}) = \textit{rank}(\mathbf{B}_0)) \to 1$ *as* $n \to \infty$.

Similar to Lasso, estimation consistency does not guarantee correct rank estimation for matrix regularization. In fact, the assumptions here are stronger than those in Theorem 2.1. For example, Theorem 2.1 allows $p + q = o(n/\log^3 n)$ while Theorem 2.2 requires $\max(p,q) = o\left(n^{1/3}\log^{-3/2} n\right)$ if $\min(p,q) = O(1)$. The proof is based on the arguments in Bach [2008a] with modifications to allow diverging p and q.

Remark 1. *Although nuclear norm penalized least squares is used to estimate the classification direction, there is a fundamental difference between our theorems and the theoretical results derived for nuclear norm penalized least squares regression [Bach, 2008a, Negahban et al., 2012]. The previous work assumes that the data obey a linear regression model with covariates-independent additive noise, which is not true in our case. In particular, the covariates* $\mathbf{X}_i$ *and the residuals*

ϵ_i are no longer independent in our problem, which brings additional challenges in developing theoretical results.

Next we state a classification error consistency result. To be consistent with the notation in the classification literature, for subject i, we use $Y_i \in \{-1, 1\}$ to denote its true label, $\hat{f}_n(\mathbf{X_i})$ as the classified label for which $\hat{f}_n$ is the classification rule obtained by solving (2.2), and $l(Y_i, f(\mathbf{X_i})) = I\{Y_i \neq \text{sign}(f(\mathbf{X_i}))\}$ as the 0-1 loss function. Define the risk of $\hat{f}_n$ as $R(\hat{f}_n) = E_{\mathbf{X}} l(Y, \hat{f}_n(\mathbf{X}))$ and the Bayes risk as $R^* = \inf_f R(f)$. In addition, we assume that the true label Y_i given $\mathbf{X_i}$ is determined by the linear classification rule with coefficients β_0^* and $\mathbf{B}_0$. Then the following theorem shows that the proposed classifier achieves the Bayes optimal risk under certain conditions. The proof, given in the Appendix B, is based on the general results in Zhang [2004], where the author studied the optimal Bayes error rate using a classifier obtained by minimizing a convex upper bound of the classification error function.

THEOREM 2.3. *Assume the same conditions for Theorem 2.1 hold and $\omega_n \to 0$. Then $R(\hat{f}_n) \to R^*$ as $n \to \infty$.*

2.4 Numerical results

2.4.1 Simulation

We conduct simulation studies to evaluate the numerical performance of our proposed method. We compare its performance with that of a few alternatives: "Lasso LDA", which adopts a naive Lasso penalty in LDA without taking into account matrix structure, the regularized matrix logistic regression [Zhou and Li, 2014a] using nuclear norm and Lasso penalties, denoted by "Logistic Nuclear" and "Logistic Lasso", and the penalized matrix discriminant analysis (PMDA) approach proposed by Zhong and Suslick [2015]. We generate $n \in \{100, 200, 500\}$ samples from two

Table 2.1: Simulation results: misclassification rates (%) and associated standard errors obtained from our method, Lasso LDA, Logistic Nuclear (L-Nuclear), Logistic Lasso (L-Lasso) and penalized matrix discriminant analysis (PMDA) based on 1000 Monte Carlo replications.

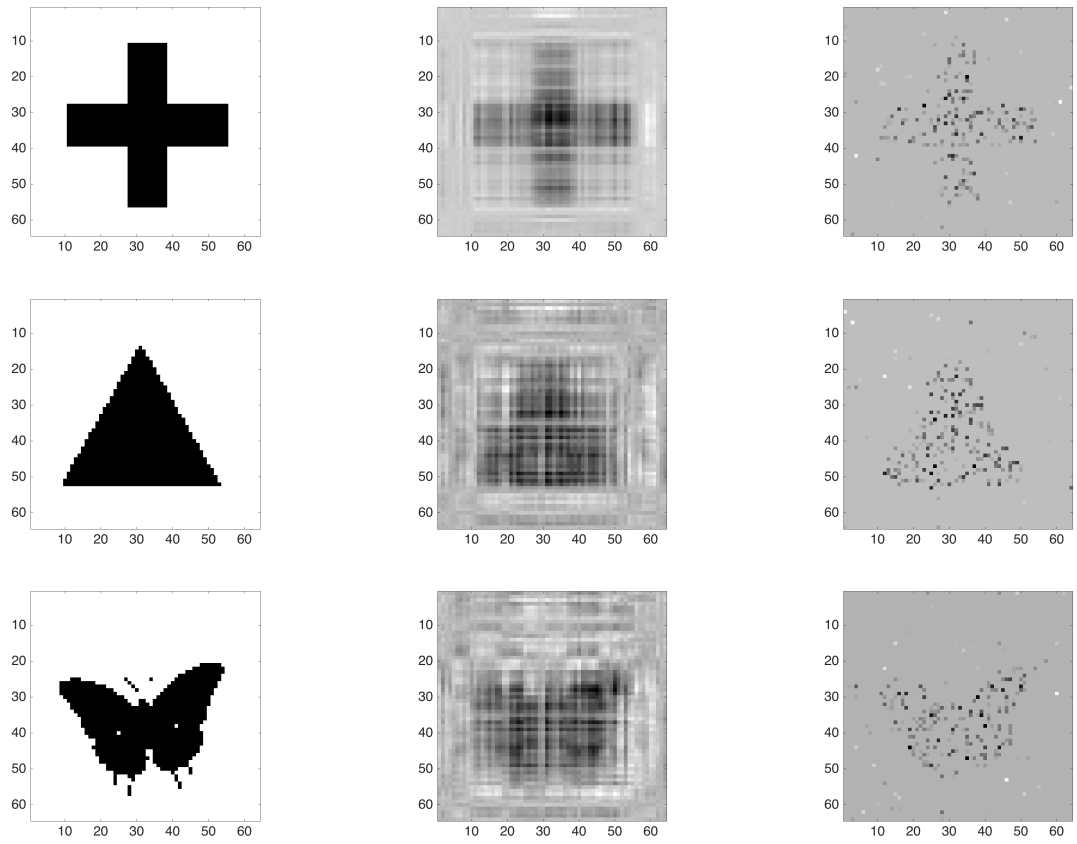
Shape	n	(π_1, π_2)	Ours	Lasso LDA	L-Nuclear	L-Lasso	PMDA
Cross	100	(0.5,0.5)	3.65(0.02)	17.81(0.07)	3.70(0.02)	19.51(0.07)	*
	100	(0.75,0.25)	3.32(0.02)	14.89(0.05)	6.62(0.04)	18.84(0.04)	*
	200	(0.5,0.5)	3.22(0.02)	11.69(0.05)	3.26(0.02)	13.39(0.05)	26.93(0.05)
	200	(0.75,0.25)	2.87(0.02)	9.89(0.04)	4.14(0.03)	16.27(0.04)	19.58(0.08)
	500	(0.5,0.5)	3.09(0.02)	6.97(0.03)	3.11(0.02)	8.19(0.04)	25.17(0.04)
	500	(0.75,0.25)	2.62(0.02)	5.81(0.03)	3.59(0.02)	14.91(0.03)	12.05(0.04)
Triangle	100	(0.5,0.5)	3.12(0.02)	15.73(0.06)	3.11(0.02)	17.70(0.07)	*
	100	(0.75,0.25)	2.66(0.02)	13.72(0.05)	6.10(0.04)	17.19(0.04)	*
	200	(0.5,0.5)	2.85(0.02)	9.90(0.04)	2.81(0.02)	11.81(0.04)	30.17(0.08)
	200	(0.75,0.25)	2.43(0.02)	8.72(0.03)	3.62(0.02)	13.40(0.04)	24.63(0.10)
	500	(0.5,0.5)	2.67(0.02)	5.67(0.03)	2.73(0.02)	6.96(0.03)	25.92(0.04)
	500	(0.75,0.25)	2.29(0.01)	4.89(0.02)	2.74(0.02)	9.97(0.03)	14.69(0.05)
Butterfly	100	(0.5,0.5)	3.86(0.02)	17.10(0.06)	4.16(0.02)	18.82(0.07)	*
	100	(0.75,0.25)	3.47(0.02)	14.79(0.05)	7.14(0.04)	17.78(0.04)	*
	200	(0.5,0.5)	3.67(0.02)	11.00(0.04)	3.78(0.02)	12.66(0.05)	29.79(0.07)
	200	(0.75,0.25)	3.26(0.02)	9.80(0.04)	4.50(0.02)	13.93(0.04)	23.83(0.09)
	500	(0.5,0.5)	3.56(0.02)	6.50(0.03)	3.52(0.02)	7.70(0.03)	25.77(0.04)
	500	(0.75,0.25)	3.02(0.02)	5.74(0.03)	3.51(0.02)	10.49(0.03)	14.66(0.05)

classes with weights $(\pi_1, \pi_2) \in \{(0.5, 0.5), (0.75, 0.25)\}$. For each class, we generate predictors from a bivariate normal distribution with means μ_g, $g = 1, 2$, and covariance $\mathbf{\Sigma}$. We set $\mu_1 = 0$ and $\mu_2 = \mathbf{\Sigma}\text{vec}(\mathbf{B}_0)$. The covariance matrix $\mathbf{\Sigma}$ has a 2D autoregressive structure: $\text{cov}(\mathbf{x}_{i_1,j_1}, \mathbf{x}_{i_2,j_2}) = 0.5^{|i_1-i_2|+|j_1-j_2|}$ for $1 \leq i_1 \leq p$ and $1 \leq j_1 \leq q$. The true signal $\mathbf{B}_0$ is generated based on a 64-by-64 image. We consider three settings: a cross, a triangle and a butterfly. These pictures are shown in Figure 2.1(a) respectively. In particular, the white color denotes value 0 and black denotes 0.05. We apply each fitted model to an independent test data set of size 1000 and summarize the misclassification rates based on 1000 Monte Carlo replications. The results are contained in Table 2.1.

The results show that our method performs much better than "Lasso LDA" and "Logistic Lasso" under all scenarios. This is expected because these two methods ignore the matrix structure. For

"Logistic Nuclear", it has similar misclassification rates with our method for balanced data, but does not perform as good as ours for unbalanced data. We have also plotted the estimates using nuclear norm and ℓ_1-norm from one randomly selected Monte Carlo replicate in Figure 2.1(b)(c). It can be seen that the proposed nuclear norm regularization is much better than ℓ_1-regularization in recovering the matrix signal in different shapes. By comparing the recovery of different shapes in Column (b) in Figure 2.1, we find that our method works better for cross than for triangle and butterfly. This is expected since triangle and butterfly do not have the low rank structure.

We also compare the performance of our method with that of PMDA proposed by Zhong and Suslick [2015]. In Table 2.1, it can be seen that our proposed method has a lower mis-classification rate under all scenarios. This is because we allow flexible values of the rank for the linear discriminant direction $\mathbf{B}$, while in Zhong and Suslick [2015], their assumption is equivalent to assuming $\mathbf{B}$ is of rank 1. In particular, using their notation, for binary case, their direction $\mathbf{B} = \boldsymbol{\beta}_1 \boldsymbol{\xi}^{\mathrm{T}}$, where $\boldsymbol{\beta}_1 \in \mathbb{R}^p$ and $\boldsymbol{\xi} \in \mathbb{R}^q$. Since the true ranks of $\mathbf{B}$ in our simulation studies are all of rank greater than 1, it is not surprising that our method outperforms PMDA. Moreover, PMDA does not apply to the case where $n < p + q$, i.e., the sample size is far smaller than the summation of image dimensions. Therefore, their method does not apply to one of our simulation settings $(n, p, q) = (100, 64, 64)$ and we mark their results using $*$ in Table 2.1. We also compare the computation time between PMDA and our method. In simulation, when $n = 200$ and true signal is a cross, given a fixed regularization parameter, the system running time of PMDA is around 1.5 minutes whereas the system running time of our method is no more than 13 seconds. Here system running time is measured on a Macbook Pro laptop with a 2.9 GHz Intel Core i5. This is because PMDA essentially solves least square problems with L_1 penalty in each iteration when setting $\omega_1 = 0$ in Algorithm 2 in Zhong and Suslick [2015]. Our method is based on the Nesterov optimal gradient method which avoids computing inverse of covariance matrix and hence has a faster convergence rate.



(a) (b) (c)

Figure 2.1: The figures for cross image: (a) original signal; (b) our nuclear regularization estimate; (c) ℓ_1-regularized estimate.

2.4.2 Real data application

We apply our method to an EEG dataset, which is available at `https://archive.ics.uci.edu/ml/datasets/EEG+Database`. The data was collected by the Neurodynamics Laboratory to study the EEG correlates of genetic predisposition to alcoholism. It contained measurements from 64 electrodes placed on each subject's scalps sampled at 256 Hz (3.9-msec epoch) for 1 second. Each subject was exposed to three stimuli: a single stimulus, two matched stimuli, two unmatched stimuli. Among the 122 subjects in the study, 77 were alcoholic individuals and 45 were controls. More details about the study can be found in Zhang et al. [1995b]. In statistics literature, EEG data has been analyzed using different models, for example, Gao et al. [2019a] con-

Table 2.2: EEG data analysis: misclassification rates (%) and associated standard errors.

Our method	Lasso LDA	Logistic Nuclear	Logistic Lasso	PMDA
22.20(0.53)	24.12(0.70)	24.44(0.80)	26.24(0.91)	*

sidered an unsupervised approach for clustering EEG data, Gao et al. [2019b] and Gao et al. [2018] considered an evolutionary state-space model and graphical model for better understanding brain connectivity, respectively. However, these methods are not directly applicable for classification purpose here.

In the data analysis, for each subject, we use the average of all 120 runs for each subject under single-stimulus condition and use that as the covariate $\mathbf{x}_i$, which is a 256×64 matrix. The classification label is *alcoholic* or not. We randomly divide the data set into training set of 81 subjects and test set of 41 subjects for 100 times, and each time fit the model on the training set and apply it on the test set to obtain the average mis-classification rate and its standard error. The results for different methods are summarized in Table 2.2. It can be seen that the proposed method has a significant lower mis-classification rate compared with other methods, which agrees with the simulation findings for the unbalanced data. PMDA does not work here since $p + q > n$ ($(n, p, q) = (122, 256, 64)$). We also check the fitted signal matrix and it agrees well with the one obtained by Zhou and Li [2014a].

In terms of computational efficiency, we measured the computation time among Lasso LDA, Logistic Nuclear, Logistic Lasso and our method based on one evaluation of the data, that is, partitioning the data into training and test sets, fitting the model on the training set and applying it on the test set. The running time for Lasso LDA, Logistic Nuclear, Logistic Lasso and our method is 0.67s, 1.79s, 1.27s and 1.87s, respectively. The system running time is measured in Matlab R2015b on a Macbook Pro laptop with a 2.9 GHz Intel Core i5.

2.5 Discussion

In the literature, total variation (TV) regularization has also been commonly used for modeling image data in addition to the proposed nuclear norm regularization. Their focuses are slightly different – the former is on structured sparse pattern and the later is on low-rank pattern. The main reason that we choose to focus on the nuclear norm regularization in this paper is because we have found that low rankness is a more reasonable assumption than sparseness assumption in our real data application. In particular, the mis-classification errors of our method is lower than the sparse method (LASSO) in our real data analysis. The TV regularization is an interesting direction to explore as it requires new computational algorithms and theories; and thus we leave this for the future research.

In this paper, we only consider the situation where all the image measurements are taking at the same scale, that is, the dimension of the image covariates p and q are equal for every study subject. We believe this is the case for most applications. For the special cases when image dimensions vary across subjects, our method may still be applicable by first resizing the image to the same scale. It will be of future interest to develop flexible statistical methods to handle image data that can be of different sizes in general.

Acknowledgment

Shen's research is partially supported by Simons Foundation Award 512620 and National Science Foundation award DMS 1509023. Kong's research is partially supported by the Natural Science and Engineering Research Council of Canada.

Chapter 3

Nonparametric Matrix Response Regression

3.1 Introduction

Large-scale neuroimaging studies have received increased attention in statistics literature, with applications including calcium imaging, electroencephalography (EEG), magnetic resonance imaging and functional magnetic resonance imaging. The obtained neuroimaging data often takes complex structure, in the form of two-dimensional matrices. For example, the EEG data can be represented as a two-dimensional matrix, where voltage values were measured from multiple electrodes placed on the subject's scalp for consecutive time points.

With recent explosive development of neuroimaging technologies, in many applications, instead of observing one 2D image, one can observe a sequence of 2D imaging data objects. It is usually of interest to model the dynamic change process of these 2D images (over the study period) and study their associations with other predictors (e.g., health information). Moreover, these image observations often possess additional structural information such as spatial/temporal correlation,

low rankness, and sparsity, which may provide useful scientific insight but also imposes additional challenges to statistical analysis. Here we discuss two relevant motivating examples. The first example is related to the fluorescent calcium imaging, which is a popular technique for observing the spiking activity of large neuronal populations. The locations of neurons and the times at which they fire can be observed via a sequence of 2D images taken over the time, typically using two-photon microscopy [Helmchen and Denk, 2005, Petersen et al., 2018]. The scientific question is to identify the major neurons and model their spiking activity over the time. As shown in Figure 3.2(a), the images are quite noisy and structured-sparse (in the sense that the signal level is low at the boundary for most images), yet contains rich information (the video clip we study in this paper composes 3,000 picture frames). Thus it is important and non-trivial to develop an automatic-yet-flexible pipeline for analyzing such type of data sets.

Our second example is the brain functional connectivity analysis. Functional connectivity refers to the coherence of the activities among distinct brain regions [Horwitz, 2003], and it provides novel insights on how distributed brain regions are functional integrated [Biswal et al., 1995, 2010, Fox et al., 2005]. Generally, studies on the functional connectivity are based on the temporal correlation between spatial remote neurophysiological events [Friston, 1994] with an implicit assumption that the functional connectivity is constant during the observation period. Recently, functional connectivity has been shown to fluctuate over time [Chang, Liu, Chen, Liu, and Duyn, 2013], implying that measures assuming stationarity over a full scan may be too simplistic to capture the full brain activity. Since the initial findings, researchers have investigated the so-called *dynamic functional connectivity*, see Calhoun et al. [2014], Calhoun and Adali [2016], Preti et al. [2017] for reviews to date. It then makes sense to represent the connectivity as a covariance matrix and model its change over the time. In our second motivating example, we analyze an EEG data set where the goal is to study the dynamic functional connectivity between alcoholic and non-alcoholic individuals. As shown in Figure 3.5, our developed methodology is capable of revealing a significant difference in terms of image pattern and temporal correlation between two groups of participants.

In this paper, we aim at developing a novel regression approach to quantify the association between 2D image outcomes and the predictors such as time, patient demographics, and other disease predictors. In particular, by including time as a predictor allows us to study the dynamic change of the images. The proposed regression model can also help detect the difference in image outcomes between study groups by including group indicator as a predictor. In contrast to the dominating use of linear models in the literature, we adopt a flexible nonparametric regression model to capture the commonly-seen nonlinear relationship in the data. For example, we performed a preliminary analysis on the calcium imaging data collected by Ilana Witten's lab at the Princeton Neuroscience Institute [Petersen et al., 2018]. Figure 3.2(b) shows a scatter plot of the changes of fluorescent intensities across time from a randomly selected pixel of the 2D-image. The scatter plot shows a clear nonlinear pattern, which will be neglected by linear models. Note that classical nonlinear regression approaches such as Nadaraya-Watson method can not be directly used in our case to model the matrix-valued image responses, since doing so is equivalent to vectorizing the 2D-image data, which destroys the underlying spatial information of the image. Instead, we maintain the matrix structure of the image data, and propose a novel low rank nonparametric estimator by solving a nuclear norm regularization problem. By the singular value thresholding algorithm [Cai et al., 2010], we show that our estimator has a closed-form solution for each fixed bandwidth and regularization parameter. To select these tuning parameters, we derive a Bayesian information criterion (BIC) based on our model and estimation procedure. For theoretical justification, we derive the risk bound for our nonparametric estimator. We show that the rank of the true function can be consistently estimated as well.

Compared with the proposed methods in the literature, here we highlight our contributions. First, we propose a novel nonparametric matrix response regression model. There are some related works on (generalized) linear models for matrix-valued data. For example, Zhou and Li [2014b] proposed a class of regularized matrix linear regression model by treating matrix data as covariates; Wang and Zhu [2017] developed a generalized scalar-on-image regression model via total variation; Ding and Cook [2018] studied the matrix response linear regression model using envelope methods;

Kong et al. [2019] proposed a low-rank linear regression model with high-dimensional matrix response and high dimensional scalar covariates. To the best of our knowledge, no work has been done on using nonparametric models for matrix data analysis. Second, our nonparametric estimator is easy to derive and has a closed-form solution, which makes it computationally more efficient than the state-of-art multivariate varying coefficient model [Zhu et al., 2011, 2012]. Third, we derive an analytic form of BIC, which is not straightforward in our nonparametric matrix response model. Finally, we develop the asymptotic theories including the risk bound and rank consistency for the proposed nonparametric estimator, which directly connects to the existing work on nonparametric statistics theory.

The rest of the article proceeds as follows. In Section 2, we introduce a novel nonparametric matrix response regression model and propose a fast algorithm for our low-rank regularized estimation procedure. We further derive a BIC for our model to choose the tuning parameters. Section 3 investigates the theoretical properties of our method. We evaluate finite performance of our method in Section 4. Section 5 illustrates applications of our method to two real datasets from a calcium imaging study and an electroencephalography study.

3.2 Method

3.2.1 Model

Suppose we observe a set of 2D-images and some scalar predictors from n independent study subjects. Let Y_i be a $p \times q$ matrix representing the 2D-image from the ith subject, and $X_i = (x_{i1}, \ldots, x_{is})^T$ be an $s \times 1$ vector denoting the scalar covariates of interest (e.g., time and disease predictors). We propose the following nonparametric matrix response model,

$$E(Y_i|X_i) = g(X_i), \tag{3.1}$$

where $g(\cdot) : \mathbb{R}^s \to \mathbb{R}^{p\times q}$ is a nonparametric matrix-valued function that quantifies the nonlinear relationship between (each pixel of) Y_i and X_i. Since $g(x)$ is a $p \times q$ matrix for all values of x, we will also impose a structure constraint on g for scientific interpretability and regularization purpose.

Our goal is to estimate the nonparametric function g. A commonly used estimator is the Nadaraya-Watson estimator for the matrix data, which can be written as

$$\hat{g}_{\text{nw}}(x) = \frac{\sum_{i=1}^n K_H(x - X_i)Y_i}{\sum_{i=1}^n K_H(x - X_i)}, \tag{3.2}$$

where $K_H(\cdot) = \frac{1}{|H|}K(H^{-1}\cdot)$, $K(\cdot)$ is a kernel function, and $H = \text{diag}(h_1, h_2, \cdots, h_s)$ is a bandwidth matrix. It is often assumed that $h_1 = \cdots = h_s = h$ for computational convenience.

However, the Nadaraya-Watson estimator is a "naive" estimator in our case because it does not utilize the underlying structure of the matrix response Y_i. In particular, the estimator in (3.2) can also be obtained by vectorizing Y_i, applying the Nadaraya-Watson estimator for the vectorized data, and transforming the estimator back to a matrix. To account for the matrix structure, we take another look at the estimator in (3.2), which can be obtained by solving the following optimization problem

$$\hat{g}_{\text{nw}}(x) = \text{argmin}_Y \sum_{i=1}^n K_H(x - X_i)\|Y_i - Y\|_F^2, \tag{3.3}$$

where $\|\cdot\|_F$ is the Frobenius norm of a matrix.

To further exploit the underlying structure of the 2D response, we introduce a penalty on Y and propose to solve

$$\hat{g}(x) = \text{argmin}_Y \left\{ \frac{1}{2n} \sum_{i=1}^n K_H(x - X_i)\|Y_i - Y\|_F^2 + \lambda_n \|Y\| \right\}, \tag{3.4}$$

where λ_n is the tuning parameter and $\|\cdot\|$ is some norm of a matrix. Possible choices are nuclear

norms, total variation norms, and their combination; and each of those norms will have different regularization effects on the image outcomes. For this paper, we mainly focus on the nuclear norm regularization for illustration, i.e., writing $\|Y\|$ as $\|Y\|_*$, which is defined as the sum of all singular values of the matrix Y. The nuclear norm is very popular in 2D-image denoising [Gu et al., 2014]. The underlying true 2D-image is often of low rank or approximately low rank, and the nuclear norm regularization can help recover the low rank structure given a noisy image [Chen et al., 2013a]. In our case, $\hat{g}(x)$ can be regarded as an image estimate at the point x, and therefore the penalty $\|Y\|_*$ can push for a low rank representation of the image estimate.

It can be shown that solving (3.4) is equivalent to solving

$$\hat{g}(x) = \text{argmin}_Y \left\{ \frac{1}{2}\|\hat{g}_{\text{nw}}(x) - Y\|_F^2 + \frac{n\lambda_n}{\sum_{i=1}^n K_H(x - X_i)}\|Y\|_* \right\}. \tag{3.5}$$

The optimization problem (3.5) can be solved using the following proposition restated from Cai et al. [2010].

Proposition 1. *Consider the singular value decomposition of a matrix $Y \in R^{p\times q}$ with rank r,*

$$Y = U\Sigma V^*, \quad \Sigma = \textit{diag}(\{\sigma_j\}_{1\le j\le r}),$$

where U and V are $p \times r$ and $q \times r$ matrices respectively with orthonormal columns, and singular values σ_j are positive. The soft-thresholding operator D_τ is defined as

$$D_\tau(Y) = UD_\tau(\Sigma)V^*, D_\tau(\Sigma) = \textit{diag}(\{(\sigma_j - \tau)_+\}_{1\le j\le r}), \tag{3.6}$$

where $(\cdot)_+$ is the positive part of $(\cdot)$. Then $D_\tau(Y)$ satisfies

$$D_\tau(Y) = \arg\min_X \left\{ \frac{1}{2}\|Y - X\|_F^2 + \tau\|X\|_* \right\},$$

where $\|X\|_*$ is defined as the nuclear norm of the matrix X.

By Proposition 1, our estimator in (3.4) can be obtained using the following algorithm.

Algorithm 1 Algorithm to solve the optimization problem (3.4).

Input: $\{(X_i, Y_i), 1 \leq i \leq n\}, x, H, \lambda_n$.

Step 1: Perform singular value decomposition of $\hat{g}_{\text{nw}}(x) = \frac{\sum_{i=1}^n K_H(x-X_i)Y_i}{\sum_{i=1}^n K_H(x-X_i)}$, and denote it by $U\Sigma V^*$. The diagonal matrix $\Sigma = \text{diag}(\{\sigma_j\}_{1\leq j\leq r})$, where $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r > 0$ with r being the rank of Σ.

Step 2: Set $\tau = \frac{n\lambda_n}{\sum_{i=1}^n K_H(x-X_i)}$, and calculate the soft-thresholding operator $D_\tau(\Sigma) = \text{diag}(\{(\sigma_j - \tau)_+\}_{1\leq j\leq r})$.

Step 3: Calculate $\hat{g}(x) = UD_\tau(\Sigma)V^*$.

Output: $\hat{g}(x)$.

3.2.2 Bayesian information criterion

The optimization problem (3.4) involves two tuning parameters, the bandwidth h and the regularization parameter λ. The choices of these two parameters are critical as they control the temporal smoothing level and the spatial low-rank level, respectively. In this paper, we derive a Bayesian information criterion (BIC) to select them. Define $\tilde{\lambda} = \frac{n\lambda_n}{\sum_{i=1}^n K_H(x-X_i)}$ and $\hat{Y}_i(\tilde{\lambda}) = \hat{g}(X_i)$. Without loss of generality, we assume $p \geq q$ and denote the singular values of $\hat{Y}_i(\tilde{\lambda})$ by $b_{i1}(\tilde{\lambda}) \geq \cdots \geq b_{iq}(\tilde{\lambda}) \geq 0$. From Algorithm 1, it can be seen that the singular values of $\hat{Y}_i(\tilde{\lambda})$ are corresponding truncated singular values of $\hat{g}_{\text{nw}}(X_i)$.

Since we are considering a least squared error loss in (3.4), the BIC can be defined as

$$\text{BIC}(\widetilde{\lambda}) = npq \log(\frac{1}{npq}\sum_{i=1}^n \|Y_i - \hat{Y}_i(\widetilde{\lambda})\|_F^2) + \log(npq)\text{df}(\widetilde{\lambda}), \tag{3.7}$$

where $\text{df}(\widetilde{\lambda})$ is given in the following proposition.

Proposition 2. *Denote $\hat{g}_{nw}(X_i)$'s singular values by $\sigma_{i1} \geq \sigma_{i2} \geq \cdots \geq \sigma_{ir_i} > 0$ and $\sigma_{ik} = 0$ for $k > r_i$. An unbiased estimator of the degree of freedom $df(\widetilde{\lambda})$ is*

$$\hat{df}(\widetilde{\lambda}) = K_H(0) \sum_{i=1}^{n} \frac{\hat{df}_i(\widetilde{\lambda})}{\sum_{j=1}^{n} K_H(X_i - X_j)}, \quad \textit{where}$$

$$\hat{df}_i(\widetilde{\lambda}) = \sum_{k=1}^{q} 1_{\{b_{ik}(\widetilde{\lambda})>0\}} \Big\{ 1 + \sum_{1 \leq j \leq p, j \neq k, k \leq r_i} \frac{\sigma_{ik}(\sigma_{ik} - \widetilde{\lambda})}{\sigma_{ik}^2 - \sigma_{ij}^2} + \sum_{1 \leq j \leq q, j \neq k, k \leq r_i} \frac{\sigma_{ik}(\sigma_{ik} - \widetilde{\lambda})}{\sigma_{ik}^2 - \sigma_{ij}^2} \Big\}. \quad (3.8)$$

Proof of Proposition 2: Efron [2004] has shown a general formula for the degree of freedom as

$$df = \sum_{i=1}^{n} \sum_{j=1}^{p} \sum_{k=1}^{q} \text{cov}(\hat{Y}_{ijk}, Y_{ijk})/\sigma^2,$$

where Y_{ijk} and $\hat{Y}_{ijk}$ are the (j,k)-th element of the Y_i and $\hat{Y}_i$, respectively. By Stein's theory of unbiased risk estimation [Stein, 1981], $\text{cov}(\hat{Y}_{ijk}, Y_{ijk}) = \sigma^2 E(\frac{\partial \hat{Y}_{ijk}}{\partial Y_{ijk}})$. Then an unbiased estimator of the degree of freedom is

$$\hat{df} = \sum_{i=1}^{n} \sum_{j=1}^{p} \sum_{k=1}^{q} \frac{\partial \hat{Y}_{ijk}}{\partial Y_{ijk}} = \sum_{i=1}^{n} \text{tr} \left(\frac{\partial \text{vec}(\hat{Y}_i)}{\partial \text{vec}(Y_i)^T} \right).$$

Note that $\hat{Y}_i(\tilde{\lambda})$ is a function of $\hat{Y}_{LSi}$, where $\hat{Y}_{LSi}$ is the usual least squared estimator of Y_i. Then we have

$$\begin{aligned} \frac{\partial \text{vec}(\hat{Y}_i(\tilde{\lambda}))}{\partial \text{vec}(Y_i)^T} &= D(\hat{Y}_i(\tilde{\lambda}))(\hat{Y}_{LSi}) \times D(\hat{Y}_{LSi})(Y_i) \\ &= D(\hat{Y}_i(\tilde{\lambda}))(\hat{Y}_{LSi}) \times \frac{K_H(X_i - X_i)}{\sum_{j=1}^{n} K_H(X_i - X_j)} I, \end{aligned}$$

where $D(A)(B) = \frac{\partial \text{vec}(A)}{\partial \text{vec}(B)^T}$ for two matrices A and B, and I is the identity matrix.

By taking the trace, we have:

$$\begin{aligned}\text{tr}\left(\frac{\partial \text{vec}(\hat{Y}_i(\tilde{\lambda}))}{\partial \text{vec}(Y_i)^T}\right) &= \text{tr}\left(D(\hat{Y}_i(\tilde{\lambda}))(\hat{Y}_{LSi})\right) \times \frac{K_H(0)}{\sum_{j=1}^n K_H(X_i - X_j)} \\ &= \hat{df}_i(\widetilde{\lambda}) \times \frac{K_H(0)}{\sum_{j=1}^n K_H(X_i - X_j)},\end{aligned}$$

where we define $\hat{df}_i(\widetilde{\lambda}) = \text{tr}\left(D(\hat{Y}_i(\tilde{\lambda}))(\hat{Y}_{LSi})\right)$. Further we assume that $\hat{Y}_{LSi}$ has distinct positive singular values $\sigma_{i1} > \sigma_{i2} > \cdots > \sigma_{ir} > 0$ and $\sigma_{ik} = 0$ for $k > r$.

By Theorem 3 in Zhou and Li [2014b], we have

$$\hat{df}_i(\widetilde{\lambda}) = \sum_{k=1}^{q} 1_{\{b_{ik}(\widetilde{\lambda})>0\}}\Big\{1 + \sum_{1\le j\le p, j\neq k, k\le r_i} \frac{\sigma_{ik}(\sigma_{ik} - \widetilde{\lambda})}{\sigma_{ik}^2 - \sigma_{ij}^2} + \sum_{1\le j\le q, j\neq k, k\le r_i} \frac{\sigma_{ik}(\sigma_{ik} - \widetilde{\lambda})}{\sigma_{ik}^2 - \sigma_{ij}^2}\Big\}. \quad (3.9)$$

Therefore,

$$\hat{df}(\widetilde{\lambda}) = K_H(0) \sum_{i=1}^{n} \frac{\hat{df}_i(\widetilde{\lambda})}{\sum_{j=1}^n K_H(X_i - X_j)}.$$

3.3 Theory

In this section, we present theoretical results of the estimation procedure in Eq. (3.4), including a risk bound of the regularized estimator and a rank consistency result. Denote the strength of regularization as λ_n and the true response as $g(X)$ given covariates X. Assume $g(X)$ has unknown rank r and denote the global minimizer of (3.4) by $\hat{g}(X)$. For any two sequences of real numbers a_n and b_n, we write $a_n \asymp b_n$ if there exists universal positive constants C_1 and C_2 such that $C_1 b_n \le a_n \le C_2 b_n$. We define $a \vee b = \max(a, b)$ and $a \wedge b = \min(a, b)$ for any $a, b \in \mathbb{R}$. With a little abuse of the notation, we use C to denote a universal constant whose value may change in different context but does not affect the results. For a matrix A and a sequence of real numbers a_n,

we write $A = O_p(a_n)$ or $A = o_p(a_n)$ if every element of A is $O_p(a_n)$ or $o_p(a_n)$.

Let $g_{jk}(x)$ be the (j,k)-th component of $g(x)$. We make the following assumptions:

Assumption 1. *We assume that* $|g_{jk}(x) - g_{jk}(y)| < C\|x-y\|^{\alpha_2}$ *with* $\alpha_2 > 0$, $1 \le j \le p, 1 \le k \le q$ *for any* $\|x-y\| < \delta$ *and some* $C > 0$, *when* $\delta > 0$ *is sufficiently small.*

Assumption 2. *Assume that* $npqh^{2\alpha_2+s} \to \infty$, $nh^{2s} \to \infty$, *and* $pqh^{2\alpha_2} \to 0$ *as* $n \to \infty$.

Assumption 3. *We assume that the kernel function* $K(\cdot)$ *is bounded on* $\mathbb{R}^s$. *In other words, there exists a constant* $k_{max} > 0$ *such that* $K(x) \le k_{\max}$ *and* $K_H(x) \le h^{-s}k_{max}$ *for any* $x \in \mathbb{R}^s$. *Moreover, we assume that there exist constants* $C_f, c_f > 0$ *such that the density function of the covariate* x *satisifes* $c_f \le f(x) \le C_f$.

Assumption 4. *Assume that* $n^{-1/2}(p \wedge q) \to 0$, $(pq)^{1/2}h^{\alpha_2}(p \wedge q)^{1/2}\lambda_n^{-1} \to 0$, $\lambda_n(p \wedge q)^2 \to 0$, *and* $\frac{pq(p \wedge q)}{n\lambda_n^2} \to 0$ *as* $n \to \infty$.

Assumption 1 assumes the α_2-smoothness for each element of the nonparametric function $g(x)$. This assumption is commonly used in multivariate function estimation literature such as Scott [2015]. It is possible to extend the results for anisotropic case in the future work based on the techniques developed in this paper. Assumptions 2 and 4 are required for estimation consistency and rank consistency. Assumption 3 is satisfied for most kernel density functions. The boundedness condition for the density funciton of x will hold if x is defined on a compact support. With these assumptions, we can state the two main theorems of this paper. Their proofs are given in the Appendix.

THEOREM 3.1. *Suppose that Assumptions 1–3 hold. We consider two cases for* p *and* q *(e.g., whether they diverge or not).*

(1) If both p *and* q *are fixed, let* $h \asymp (\frac{\log n}{n})^{\frac{1}{2\alpha_2+s} \wedge \frac{1}{2s}}$ *and* $\lambda_n \asymp h^{\alpha_2}$, *then*

$$\|\hat{g}(x) - g(x)\|_F^2 \le Cr\left(\frac{\log n}{n}\right)^{\frac{2\alpha_2}{2\alpha_2+s} \wedge \frac{\alpha_2}{s}}.$$

(2) If $p \vee q \to \infty$ *and* $p \vee q = o\left(n^{\frac{\alpha_2}{2\alpha_2+s}} \wedge (\frac{n}{\log n})^{\frac{\alpha_2}{2s}}\right)$*, then by letting* $h \asymp n^{-\frac{1}{2\alpha_2+s}} \vee (\frac{\log n}{n})^{\frac{1}{2s}}$ *and* $\lambda_n = h^{\alpha_2}(pq)^{1/2}$*, we have*

$$\|\hat{g}(x) - g(x)\|_F^2 \leq Cpqr\left(n^{-\frac{2\alpha_2}{2\alpha_2+s}} \vee (\frac{\log n}{n})^{\frac{\alpha_2}{s}}\right).$$

Note that the risk bound involves two quantities $n^{-\frac{2\alpha_2}{2\alpha_2+s}}$ and $(\log n/n)^{-\frac{\alpha_2}{s}}$. As the number of predictors s increases, it becomes more difficult to estimate $g(x)$. Meanwhile, h^s is involved when proving strongly restricted convexity of the loss function. A larger value of s indicates smaller probability of the loss function being strong restricted convex. In contrast, α_2 describes the smoothness of $g(x)$. A larger α_2 leads to a smaller risk bound and a faster convergence rate.

Remark 2. *If* p *and* q *are fixed and* $s \leq 2\alpha_2$*, the optimal bandwidth* h *can be chosen arbitrarily close to* $n^{-\frac{1}{2\alpha_2+s}}$*, which leads to the same convergence rate (with additional logarithmic factor) for estimating an* α_2*-smooth,* s*-dimensional function without regularization.*

Remark 3. *When* $\max(p,q) \to \infty$*, if we further assume* $s \leq \alpha_2$ *and choose* $nh^{2\alpha_2+s} \asymp \frac{(\sqrt{p}+\sqrt{q})^2}{pq}$ *and* $\lambda_n \asymp (pq)^{1/2}\left(\frac{(\sqrt{p}+\sqrt{q})^2}{npq}\right)^{\frac{\alpha_2}{2\alpha_2+s}}$*, as we let* $\lambda_n \to 0$ *and* $nh^{2s} \to 0$*, we obtain* $\max(p,q) = o(n^{\frac{\alpha_2}{\alpha_2+s}})$*. This is the necessary condition for* $\hat{g}(x)$ *being consistent. The assumption* $s \leq \alpha_2$ *rules out the case where there are too many covariates in the model.*

Next we present the rank consistency result. We consider three general cases for different values of p and q (e.g., whether they diverge or not) and discuss the corresponding choices of λ_n and h as folllows,

(C1) If both p, q are fixed, we can choose $h \asymp (\frac{\log n}{n})^{\frac{1}{2\alpha_2+s} \wedge \frac{1}{2s}}$ and $\lambda_n \asymp h^{\alpha_2} \log n$.

(C2) If $p \wedge q$ is finite, and $p \vee q \to \infty$ satisfying $(\log n)^2(p \vee q) = o((\frac{n}{\log n})^{\frac{\alpha_2}{s}} \wedge n^{\frac{2\alpha_2}{2\alpha_2+s}})$, then we choose $h = (\frac{\log n}{n})^{\frac{1}{2s}} \vee n^{-\frac{1}{2\alpha_2+s}}$ and $\lambda_n \asymp (p \vee q)^{1/2} h^{\alpha_2} \log n$.

(C3) If $p \asymp q$, and $p \to \infty$, then we let $h \asymp (\frac{\log n}{n})^{\frac{1}{2s}} \vee n^{-\frac{1}{2\alpha_2+s}}$ and $\lambda_n \asymp p^{\frac{3}{2}} h^{\alpha_2}(\log n)$. In addition, we assume $(\log n)^{\frac{2}{7}} p = o(h^{-2\alpha_2/7})$.

THEOREM 3.2. *Suppose that Assumptions 1–4 hold, and one of the cases in (C1)–(C3) holds, then* $\hat{g}(x)$ *is consistent and rank consistent, i.e.,*

$$P\{\textit{rank}(\hat{g}(x)) = \textit{rank}(g(x))\} \to 1, \quad \textit{as } n \to \infty.$$

It can be seen that rank consistency requires stronger assumptions on p, q compared with those in Theorem 3.1. For instance, the desired λ_n is much larger than the one from the previous theorem. Meanwhile, pq is not allowed to be greater than n for rank consistency.

3.4 Simulation

In this section, we evaluate the performance of our method and other competing methods. We consider both univariate and multivariate X, different nonparametric functions and different correlation structures of the random error E_i, where $E_i = Y_i - g(x_i)$.

3.4.1 Univariate predictor

Setting I: We set the dimensions of the image $p = q = 64$, and set the (j,k)-th element of the nonparametric function $g(x)_{jk} = \{\sin(10\pi x) + \cos(10\pi x) + 0.1(j+k)\} * B_{jk}$, $1 \le j, k \le 64$, where $0 \le x \le 1$ and B_{jk} is the (j,k)-th element of the true signal B. The true signal B is generated from a 64-by-64 image, where we consider three shapes: a cross, a square and a T-shape. We have plotted the true shapes in Figure 1(a)(b)(c), where we assign B a value of 5 for black regions and 0 for white regions. The sample size is set at $n = 200, 500$. The covariates $\{x_i\}, i = 1, 2, \ldots, n$ are equally spaced on [0,1]. The response Y_i is generated from $Y_i = g(x_i) + E_i$, where vec(E_i)'s are

i.i.d $N(0, I_{pq})$. The optimal bandwidth h and λ are selected by BIC. For the kernel function, we use the standard gaussian kernel defined as $K(x) = \exp(-x^2/2)/\sqrt{2\pi}$. We compare our method with the naive Nadaraya-Watson estimator and the Lasso estimator, where the Lasso estimator is obtained by solving the following optimization problem

$$\hat{g}_{\text{lasso}}(x) = \text{argmin}_Y \left\{ \frac{1}{2n} \sum_{i=1}^{n} K_H(x - X_i) \|Y_i - Y\|_F^2 + \lambda_n \|Y\|_1 \right\}, \tag{3.10}$$

where $\|Y\|_1$ is defined as the sum of the absolute values of all the elements of the matrix Y. We also use the BIC as defined in (3.7) to choose the tuning parameter for Lasso. Here the degree of freedom can be obtained by the chain rule as

$$\begin{aligned} \hat{\text{df}} &= \sum_{i=1}^{n} \text{tr}\left(\frac{\partial \text{vec}(\hat{Y}_i)}{\partial \text{vec}(Y_i)} \right) = \sum_{i=1}^{n} \sum_{j=1}^{p} \sum_{k=1}^{q} \frac{\partial \hat{Y}_{ijk}}{\partial \hat{g}_{ijk(nw)}} \frac{\partial \hat{g}_{ijk(nw)}}{\partial Y_{ijk}} \\ &= \sum_{i=1}^{n} \frac{K_H(0)}{\sum_{j=1}^{n} K_H(X_i - X_j)} \|\text{sign}(\text{vec}(\hat{Y}_i))\|_1, \end{aligned}$$

where we have used the fact that $\frac{\partial \hat{Y}_{ijk}}{\partial \hat{g}_{ijk(nw)}} = |\text{sign}(\hat{Y}_{ijk})|$.

In each Monte Carlo simulation, we generate n samples as the training set and another 500 samples as the test set. We report the integrated error $\int_x \|\hat{Y}(x) - Y(x)\|_F^2 dx$, which can be approximated by $\frac{1}{500} \sum_{i=1}^{500} \|\hat{Y}(x_i^{test}) - Y(x_i^{test})\|_F^2$. Table 3.1 shows the average integrated test error by our method, naive Nadaraya-Watson estimator and Lasso estimator based on 100 Monte Carlo replicates. We also report the average selected rank by our method using BIC, defined as $\frac{1}{n} \sum_{i=1}^{n} \text{rank}(\hat{Y}(x_i))$.

From the results, we can see that our method performs better than Nadaraya-Watson estimator and Lasso estimator in all cases. In addition, our method can estimate the true rank of the image accurately. We have plotted the recovered signals from one randomly selected Monte Carlo study in Figure 3.1(d)(e)(f), and our method manages to recover the true signals very well.

Setting II: In this setting, we consider the case where the errors $\text{vec}(E_i)$ are correlated across

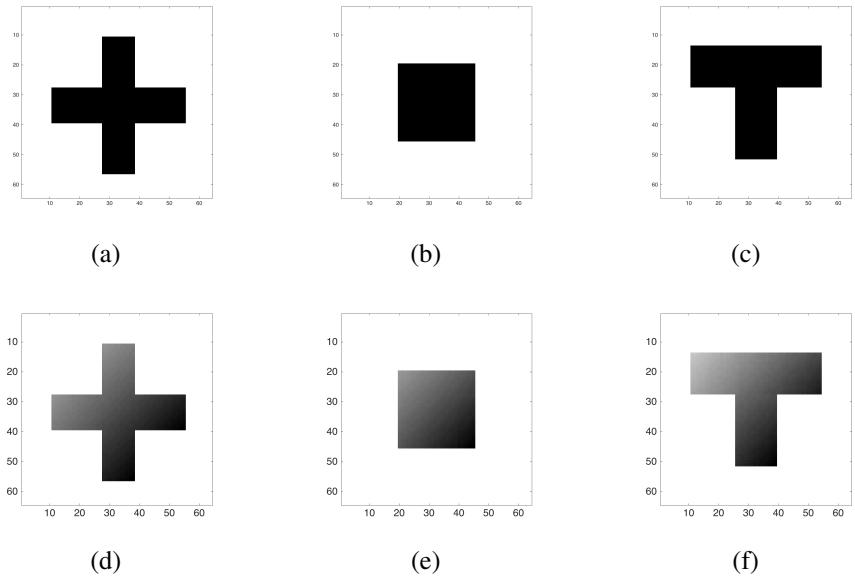


Figure 3.1: (a)-(c): true signals, (d)-(f): recovered signals

Table 3.1: Simulation results for Setting I: mean of integrated test error and associated standard errors obtained from our method, NW estimator, and Lasso, the average selected rank and true rank are reported for three different shapes B. The results are based on 100 Monte Carlo replications.

	Shape	Our method	NW	Lasso	Selected rank	True rank
n = 200	Cross	4458 (0.70)	6024 (0.96)	5148 (0.81)	3.53 (0.004)	4
	Square	4288 (0.82)	6107 (0.99)	4875 (0.83)	2.00 (0.000)	2
	Tshape	4472 (0.76)	6094 (0.98)	5193 (0.92)	3.22 (0.006)	4
n = 500	Cross	4306 (0.41)	5009 (0.52)	4560 (0.48)	3.99 (0.000)	4
	Square	4186 (0.41)	4803 (0.48)	4440 (0.45)	2.01 (0.000)	2
	Tshape	4255 (0.60)	5042 (0.51)	4579 (0.49)	3.52 (0.007)	4

Table 3.2: Simulation results for Setting II: mean of integrated test error and associated standard errors obtained from our method, NW estimator, and Lasso, the average selected rank and true rank are reported for three different shapes B. The results are based on 100 Monte Carlo replications.

	Shape	Our method	NW	Lasso	Selected rank	True rank
n = 200	Cross	4656 (2.23)	6120 (3.47)	5528 (3.49)	6.64 (0.009)	4
	Square	4463 (2.31)	5785 (3.27)	5169 (4.72)	4.42 (0.009)	2
	Tshape	4667 (2.49)	6017 (3.65)	5591 (3.74)	6.52 (0.008)	4
n = 500	Cross	4403 (1.23)	5240 (1.96)	4769 (1.71)	6.90 (0.008)	4
	Square	4296 (1.10)	5036 (1.64)	4692 (1.68)	4.76 (0.000)	2
	Tshape	4404 (1.21)	5057 (1.60)	4797 (1.69)	6.75 (0.008)	4

different subjects i's and the pixels within the same random error matrix E_i are also correlated. Define $\mathbf{e} = (\text{vec}(E_i)^{\text{T}}, \ldots, \text{vec}(E_n)^{\text{T}})^{\text{T}} \in \mathbb{R}^{pqn}$. We assume $\mathbf{e} \sim N(0, \Sigma)$, where $\Sigma = \Sigma_1 \otimes \Sigma_2 \in \mathbb{R}^{pqn \times pqn}$. Here Σ_1 is a $n \times n$ matrix representing the correlation within different subjects $1 \leq i \leq n$, Σ_2 is a $pq \times pq$ matrix representing the correlation among different pixels of the 2D image, and $\otimes$ is the Kronecker product. This decomposition of Σ is often referred to as the separability of the covariance matrix, which was studied in various literatures such as De Munck et al. [2002], Dawid [1981]. For Σ_1, we assume it has a subject-wise 1D autoregressive structure. In particular, we set the (i_1, i_2)-th element of Σ_1 as $0.5^{|i_1 - i_2|}$ for $1 \leq i_1, i_2 \leq n$. For Σ_2, we assume it is incorporated with a pixel-wise 2D autoregressive structure. Specifically, we set the $(j_1 + (k_1 - 1)q, j_2 + (k_2 - 1)q)$-th element of Σ_2 as $0.5^{|j_1 - j_2| + |k_1 - k_2|}$ for $1 \leq j_1, j_2 \leq p$ and $1 \leq k_1, k_2 \leq q$. The average integrated test errors by three methods and the average selected rank of our method are summarized in Table 3.2. From the results, we can see that our methods still outperforms than Nadaraya-Watson estimator and Lasso estimator in all cases. Compared with independent error case, one notice that we may over select the rank a bit, possibly due to the error correlations, however, the average integrated errors are still similar for both cases.

3.4.2 Multivariate predictors

Setting III: We consider shapes of the image with the same pixel value as setting I. We set the

(j,k)-th element of the nonparametric function $g(x)_{jk} = \{\sin(2\pi\|x\|) + \cos(2\pi\|x\|) + 0.5(j+k)\} * B_{jk}$, $x \in [0,1] \times [0,1]$, $1 \leq j,k \leq 64$, where we consider the same three shapes of the true image B and $\|x\|$ is the l_2-norm of x. The random error vec(E_i)'s are i.i.d. $N(0, I_{pq})$. The covariates x_i consist of a set of $\{x_{jk}\}$, $1 \leq j \leq 20, 1 \leq k \leq 25$, that are equally spaced on $[0,1] \times [0,1]$. The sample sizes $n = 200, 500$ are considered, and the multivariate Gaussian kernel defined as $K(x) = \exp(-\|x\|^2/2)/2\pi$ is used. In each Monte Carlo simulation, we generate n samples as the training set and another 500 samples as the test set. We report the average integrated test error obtained by our method, the naive Nadaraya-Watson estimator and Lasso estimator and the average selected rank of our method based on 100 Monte Carlo replicates in Table 3.3.

Table 3.3: Simulation results for Setting III: mean of integrated test error and associated standard errors obtained from our method, NW estimator, and Lasso, the average selected rank and true rank are reported for three different shapes B. The results are based on 100 Monte Carlo replications.

	Shape	Our method	NW	Lasso	Selected rank	True rank
n = 200	Cross	4711 (0.86)	7021(1.17)	5759(1.01)	4.03 (0.002)	4
	Square	4518 (0.80)	6723(1.09)	5326(1.05)	2.04 (0.002)	2
	Tshape	4719 (0.83)	7137 (1.11)	5829(1.15)	4.34 (0.005)	4
n = 500	Cross	4647 (0.49)	5562 (0.62)	4843 (0.48)	4.84 (0.006)	4
	Square	4281 (0.42)	5506 (0.58)	4649(0.45)	2.01 (0.001)	2
	Tshape	4376 (0.45)	5620 (0.59)	4875 (0.49)	4.12 (0.002)	4

Setting IV: We consider the same setting as Setting III except that the random error vec(E_i)'s are correlated across different i's and the pixels within the same random error matrix E_i are also correlated. The random error vec(E_i)s are generated the same as Setting II. The average integrated test errors by three methods and the average selected rank of our method are summarized in Table 3.4.

The findings in the multivariate case (Settings III and IV) are consistent with the ones in the univariate case. The simulation results in this section confirm the excellent performance of the proposed nonparametric estimation procedure.

Table 3.4: Simulation results for Setting IV: mean of integrated test error and associated standard errors obtained from our method, NW estimator, and Lasso, the average selected rank and true rank are reported for three different shapes B. The results are based on 100 Monte Carlo replications.

	Shape	Our method	NW	Lasso	Selected rank	True rank
n = 200	Cross	4894 (2.58)	7164 (3.89)	5804 (4.13)	5.40 (0.008)	4
	Square	4642 (2.43)	6858 (3.57)	5360 (4.0)	3.14 (0.009)	2
	Tshape	4910 (2.66)	7283 (3.91)	5880 (4.36)	5.36 (0.008)	4
n = 500	Cross	4779 (1.73)	5687 (2.38)	5067 (1.86)	6.38 (0.008)	4
	Square	4574 (1.51)	5614 (2.22)	4815 (1.62)	4.12 (0.001)	2
	Tshape	4797 (1.55)	5745 (2.09)	5080 (4.72)	6.34 (0.008)	4

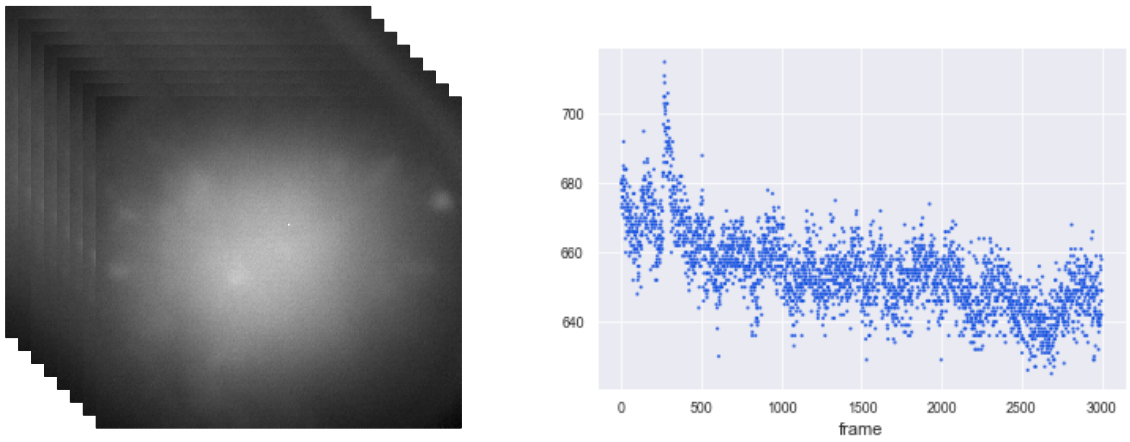
3.5 Real data application

3.5.1 Application to calcium imaging data study

In this section, we apply the proposed method to one-photon calcium imaging dataset collected by Ilana Witten's lab at the Princeton Neuroscience Institute [Petersen et al., 2018], which can be downloaded from `https://ajpete.com/software`. Calcium imaging is an important fluorescent microscopy technique regulating a great variety of neuronal processes simultaneously [Berridge, 1998, Andilla and Hamprecht, 2014]. Whenever a neuron fires, voltage-gated calcium channels in the axon terminal open and then calcium floods the cell. Such changes in concentration of calcium ions are detected by observing the fluorescence of calcium indicator molecules. Therefore, not surprisingly, intracellular calcium concentration becomes an important surrogate marker for the spiking activity of neurons in the absence of effective voltage imaging approach and is commonly used when analyzing local neuronal circuits in vivo and in vitro [Petersen et al., 2018, Grienberger and Konnerth, 2012].

The calcium imaging data can be viewed as a video clip (i.e., a collection of 2D-images recorded at the same frame over a period of time) that presents the location and time of neuron firing [Apthorpe et al., 2016, Petersen et al., 2018]. Each pixel in a frame is continuous-valued and larger values in-

dicate higher fluorescent intensities caused by greater calcium concentrations. The calcium imaging video we used consists of 3000 frames of size 205×226 pixels sampled at 10 Hz. An example frame randomly selected from the video is shown in Figure 3.4(a).



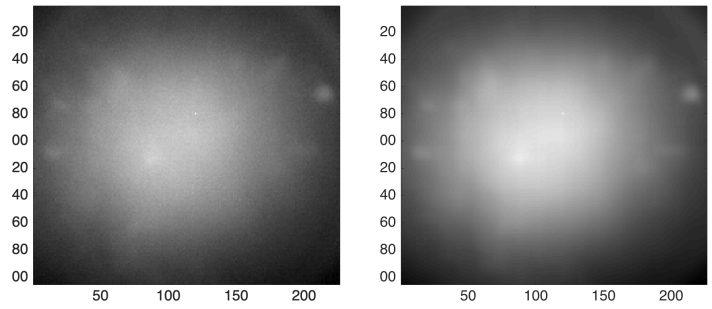
(a) (b)

Figure 3.2: (a) A sequence of frames (b) scatterplot for a fixed voxel of coordinate (200,60) over frames

(a) (b)

Figure 3.3: (a) Fitted value for a fixed voxel of coordinate (200,60) over frames (b) Fitted value for a fixed voxel of coordinate (60,180) over frames

Figure 3.3 gives estimated fluorescent intensities versus true values of two randomly selected pixels across 3000 frames. The optimal bandwidth and regularization parameter are selected by the proposed BIC criteria. The average rank selected by our method is $22.34(\text{SE} = 1.34)$ across all frames. We also plot estimated images from a randomly selected frame by our method in Figure



(a) (b)

Figure 3.4: (a) Original 1500th frame (b) Estimated 1500th frame by our method

3.4(b). From that figure, we can see that our method amplifies potential neuron signals, but also weakens those unclear and smaller neurons.

We further evaluate the prediction performance of our method by cross-validation. We compare our method with two nonparametric regression methods: Nadaraya-Watson regression and Lasso defined in (3.3) and (3.10), respectively. We also compare our method with the low-rank matrix response linear regression (L2RM) method [Kong et al., 2019]. As shown in Table 3.5, our method reaches the smallest leave-one-out cross-validation error among four methods. For Lasso estimator, the selected tuning parameter is always zero, hence making the Lasso estimator equivalent to the Nadaraya-Watson estimator. This confirms that the sparsity assumption does not seem plausible for the calcium imaging application, while low rankness is a more reasonable assumption. The linear model L2RM has the highest prediction error, which validates the use of a nonparametric model for the data set.

Table 3.5: Leave-one-out cross-validation errors (Err)(10^6) and associated standard deviation by three methods for calcium imaging data

Err - Our method	Err - NW	Err - Lasso	Err - L2RM
$2.66(0.63)$	$2.91(1.04)$	$2.91(1.04)$	$5.45(3.63)$

3.5.2 Application to EEG data study

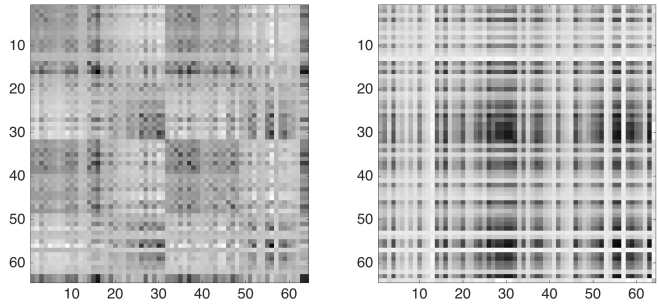
We also apply our method to an EEG dataset, which is available at `https://archive.ics.uci.edu/ml/datasets/EEG+Database`. The data was collected from 122 subjects by the Neurodynamics Laboratory to examine the EEG correlates of genetic predisposition to alcoholism. More details about the study can be found in Zhang et al. [1995a]. Among the 122 subjects, 77 were alcoholic individuals and 45 were controls. The dataset included voltage values from 64 electrodes placed on each subject's scalps sampled at 256 Hz (3.9- msec epoch) for 1 second. Each subject was exposed to three stimuli: a single stimulus, two matched stimuli, two unmatched stimuli. For each subject, we use the average of all trials for each subject under single-stimulus condition, which results in a 256×64 matrix. Among those 122 subjects, we randomly select one alcoholic individual and one control, and analyze the dynamic functional connectivity among different electrodes across time. The simplest analytical strategy to investigate dynamic functional connectivity consists in segmenting the time courses from spatial locations into a set of temporal windows, inside which their pairwise connectivity is probed. By gathering functional connectivity descriptive measures over subsequent windows, fluctuations in connectivity can be captured. The basic sliding window framework has been applied by the neuroimaging community to understand how brain dynamics related to our cognitive abilities [Kucyi and Davis, 2014, Elton and Gao, 2015, Madhyastha and Grabowski, 2014], is affected by brain disorders [Sakoğlu et al., 2010, Jones et al., 2012], or compares to other functional or structural brain measures [Leonardi et al., 2013, Tagliazucchi et al., 2012, Liégeois et al., 2016]. More specifically, we use a moving window of size 100 to calculate a series of covariance matrices along dimension of 256, resulting 157 covariance matrices of size 64×64 for each individual.

We apply the proposed method to analyze the dynamic change of covariance structures over the time in both alcoholic individual and control. The optimal bandwidth and regularization parameter are selected by BIC. Figure 3.5 shows estimated images of 10th frame by our method for alcoholic individual and control respectively. We observe a significant structural difference in their covari-

ance matrices. Specifically, the alcoholic individual has a more complex covariance structure than that from the control. Moreover, the average selected rank of alcoholic individual is 22.44 (SE = 0.93) compared to 6.82 (SE = 0.87) of control. This can be explained by drastic fluctuation across time in EEG signals of alcoholic individuals compared to stable variation in control. We further evaluate our method by leave-one-out cross-validation error and compare it with Nadaraya-Watson regression, Lasso and L2RM. As shown in Table 3.6, our method achieves the smallest leave-one-out cross-validation error among three methods. For Lasso estimator, the selected tuning parameter is always zero. In other words, the Lasso estimator is the same as the Nadaraya-Watson estimator for this data application, which implies that the low rankness assumption is a more reasonable assumption than sparsity. We also notice that linear L2RM has a much higher estimation error than the nonparametric methods. This indicates a strong nonlinear pattern in EEG signals for both alcoholic and control subjects.

Table 3.6: Leave-one-out cross-validation errors (SE) by three methods for EEG data

	Err - Our method	Err - NW	Err - Lasso	Err - L2RM
Alcoholic	1.05(1.20)	2.20(2.84)	2.20(2.84)	908.29(623.54)
Control	9.57(61.44)	21.87(118.06)	21.87(0.75)	22513.17(16763.92)



(a) (b)

Figure 3.5: (a) Estimated 10th frame for alcoholic (b) Estimated 10th frame for control

Chapter 4

Latent Representer Values in Image Classification

4.1 Introduction

The last several years have witnessed the rise of sophisticated machine learning systems on a variety of challenging tasks such as image classification or localization [Krizhevsky et al., 2012, He et al., 2016] natural language processing or speech recognition [Chorowski et al., 2015] and medical diagnosis [Choi et al., 2016]. For these settings, neural networks are the core toolkits of human decision making pipelines. Currently, machine learning models are evaluated by a variety of accuracy metrics on an available validation. Although there have been considerable improvements in the state of the art of system accuracy, even on par with human performance, unfortunately, predictions from these large neural models seem hard to interpret and be poorly calibrated. Techniques for understanding and interpreting why the neural network make this prediction therefore become an essential component of a robust validation procedure. Interpretability is gaining increasing attention in applications, e.g., data-driven precision medicine, where understanding the contribution

of specific features to a model is substantial. It is very important to understand the decision making systems before the decision from machine is accepted in clinical procedure since it closely affects the life and mortality of a patient [Katuwal and Chen, 2016]. Another application is self-driving car, where people are interested in highlighting most salient image sections by which the model makes the prediction of steering angles [Bojarski et al., 2017].

There are numerous avenues of work on interpreting black-box models but focusing on understanding how an individual prediction is made from a model, e.g., approximating the prediction locally with a probably simpler, and interpretable model [Ribeiro et al., 2016], or perturbing the point by an infinitesimal mass to study the influence on the prediction [Koh and Liang, 2017]. However, to our best of knowledge, none of them inspects the importance of each training point has on the prediction of the model, with one exception. Yeh et al. employ a representer theorem that can be generalized for deep neural networks. They show that the pre-activation prediction values can be decomposed into a sum of weights, called representer values, of training points, to measure the importance of each training point on the pre-activation value of a given test point. For training points with significant representer values, we could say these are influential points to model prediction, aiding users in which instance to inspect, and further in understanding of model's prediction. Yeh et al. measure the importance of training points by computing representer values from complex, high-dimensional data, however, we believe this would amplify or dilute the importance in terms of the ranking of representer values due to complicated data distribution in real life.

Our work proposes to tackle with this question by formalizing the impact of training points in a continuous, low-dimensional latent space considering that the latent representations can capture semantic variation of the data. We show that there still holds a representer theorem which decomposes the pre-activation value into a sum of representer values of training points. In contrast to Yeh et al.'s work, we believe that the representer values computed from latent space, characterize the "true" influence of each training point on the prediction in the sense of approximating the true ranking of influences of training data. Then we can dive into training points with significant positive

or negative representer values and understand why model tends to make the prediction of a certain class rather than other classes. For training point with positive representer value for a certain class, it indicates that a similarity to the training point encourages the prediction of the test point towards this class. For the one with negative representer value, similarity to this training point is inhibitory. We demonstrate that our method can select truly influential points through a range of theoretical properties including ranking consistency and empirical experiments: debugging dataset; detecting dataset errors; manifold visualization.

4.2 Related Work

There are several main approaches to interpreting model predictions. The first class of approach is feature based. One approach proposed by Baehrens et al. is to explain the local decision by local explanation vectors in the classification setting. The local explanation vector is in spirit to sensitivity analysis, a local gradient that describes the influence of moving a single data point locally on its predicted label. The local explanation can be used to extract important features since it can answer which local direction is influential to the prediction. Ribeiro et al. introduce model-agnostic interpretation based on if-then rules, called Anchors, which sufficiently nail down the prediction locally such that it is invariant under changes to the rest of the features. In other words, Anchors highlight important features that are sufficient for the classifier to make a certain prediction. Other gradient based methods include Pixel-space gradient visualizations such as Guided Backpropagation [Springenberg et al., 2014] and Deconvolution [Zeiler and Fergus, 2014], however, neither of them is class-discriminative. Selvaraju et al. propose Gradient-weighted Class Activation Mapping (Grad-CAM), a generalization of CAM [Zhou et al., 2016]. Grad-CAM is a class-discriminative localization technique that computes the gradient information flowing into the feature maps of the CNN to understand the importance of each neuron with particular classes. Lundberg and Lee present a unified framework, SHAP (SHapley Additive exPlanations), to explain

model prediction by assigning each feature an importance value for a certain prediction. SHAP values provide a unified measure of feature importance approximated from current additive feature attribution methods including LIME, DeepLIFT, Layer-Wise Relevance Propagation and Classic Shapley Value Estimation [Ribeiro et al., 2016, Shrikumar et al., 2017, Bach et al., 2015, Datta et al., 2016, Lipovetsky and Conklin, 2001]

The second class of approach is perturbation-based forward propagation. Inspired by influence function [Hampel, 1986], a versatile technique from robust statistics measuring the effect of an infinitesimal contamination at a single point on the estimate, Koh and Liang apply it to understanding black-box model behaviors in terms of how model parameter changes by upweighting or perturbing at a training input from the sample space. Using influence functions to inspect the training points allows people to perform data debugging such as fixing mislabeled instances. Moreover, influence function can be used to generate adversarial training images fool the model, by perturbing a training points to increase the loss on a given test point. Sharchilev et al. apply influence function to tree ensemble-based models such as Random Forest(RF) and Gradient Boosted Decision Trees(GBDT) to find influential training samples. Cadamuro et al. consider the task of identifying the small subset of training items, which are root cause of biasing the model towards creating prediction error. In addition, prediction error can be at least mitigated if this small subset is fixed. However, they only provide closed form solution for OLS and Gaussian processes.

Yeh et al. propose a representer theorem for deep neural network predictions. Representer theorem origins from Kimeldorf and Wahba and is generalized by Schölkopf et al., where the minimizer of an empirical risk with monotonical regularization term in a reproducing kernel Hilbert space(RKHS) can be expressed as expansions in terms of the training examples under certain conditions. In Yeh et al.'s work, they consider a framework of neural network explanation, under which they decompose preactivation neural network prediction into representer values of training points. Then training points with positive representer values are interpreted as excitatory examples and those with negative values as inhibitory instances. There is not too much other literature of

representer theorem with application to deep neural network, with some exceptions [Unser, 2018, Bohn et al., 2019]. However, neither of these works can interpret model prediction. Our approach extends Yeh et al.'s work towards latent space, considering that the low-dimensional latent representation of high-dimensional data can capture semantic variation in the data distribution. There are two main frameworks of learning feature representations in terms of modeling how the data are generated from the joint distribution of observed and target variables: Variational Autoencoder (VAE) [Kingma and Welling, 2013] and Generative Adversarial Network (GAN) [Goodfellow et al., 2014]. Donahue et al. present Bidirectional Generative Adversarial Networks (BiGANs) to add an inverse mapping in addition to discriminator and generator, projecting data back into low-dimensional latent space. We use BiGANs to encode data into the latent space and learn representer values of the latent representations.

4.3 Framework

In this section, we describe the problem setup and theoretical details of our framework for representer point selection on latent space. Consider a classification problem, given training dataset $x_1, x_2, \cdots, x_n$ from a corpus of data X and outputs of labels $y_1, y_2, \cdots, y_n$ from a corpus of labels Y, it is of interest to learn a mapping from X to Y. To learn the mapping, we deploy a neural network having the form $\hat{y}_i = \sigma(\phi(x_i, \theta) \in \mathbb{R}^K$, where K is the number of classes. Furthermore, $\phi(x_i, \theta) = \theta_1 f_i \in \mathbb{R}^K$ and $f_i = \phi_2(x_i, \theta_2)$.

Our goal is to quantify to what extent does each training point x_i contribute to the prediction y_t of a test point x_t. However, in general, an instance x_t may not be in X, instead, comes from the same underlying distribution P_X. Unlike the existing approach [Yeh et al., 2018] that focuses on the complicated input space, we want to learn the contributions from the manifold that captures the semantic variation in data distribution P_X, rather than from the raw data representation. This motivates us to learn the mapping from corresponding dense representation of Z space to Y di-

rectly. More concretely, we first represent instance x into dense vector z underlying latent space Z which defines the distribution P_X. Therefore powerful deep generative models are needed to learn a mapping from dense latent representation Z to P_X, from which samples $\{x_i\}_{i=1}^n$ are generated.

To achieve this, we consider Generative Adversarial Networks(GANs), a powerful class of generative models which can be trained to learn arbitrarily complex data distributions via an adversarial process between two competing networks: generator and discriminator [Goodfellow et al., 2014]. The generator G learns to map samples from an arbitrary low dimensional noise distribution to data and the adversarial discriminator D is trained to distinguish synthetic data from real samples. However, GANs can not be directly used to learn latent representations for an arbitrary data distribution, i.e., an inverse mapping from data to latent representation is lacking in this framework. BiGAN [Dumoulin et al., 2016] propose an encoder E besides the generator G and discriminator D, mapping data x to latent representations z. Moreover, the BiGAN discriminator D is also modified to discriminate jointly in pairs $(x, E(x))$ versus $(G(z), z)$ as shown in Figure 4.1. Donahue et al. refine the minimax objective as $\min_{G,E} \max_D V(D, E, G)$ where

$$\begin{aligned} V(D,E,G) =& E_{x\sim P_X}\left[E_{z\sim E(\cdot|x)}[\log D(x,z)]\right] \\ &+ E_{z\sim P_Z}\left[E_{x\sim G(\cdot|z)}[1-\log D(x,z)]\right]. \end{aligned}$$

Donahue et al. proves that BiGANs keep many of the theoretical properties of GANs while the encoder E and generator G are able to learn to invert each other to fool the BiGAN discriminator D. Therefore we consider a trained BiGAN encoder as a powerful and robust tool to extract latent representation for semantic tasks.

Then let $L(z, y, \theta)$ be the loss and our objective is to minimize empirical loss with a regularization term $\frac{1}{n}\sum_{i=1}^n L(z_i, y_i, \theta) + \|\theta_1\|_2$ where $\|\cdot\|_2$ is l_2 norm. Suppose the optimal solution is θ^*, it would be ideal if $\phi(z_t, \theta^*) = \sum_{i=1}^n \alpha_i k(z_t, z_i)$, i.e., $\phi(z_t, \theta^*)$ has a decomposition as a sum of n weights where $\alpha_i k(z_t, z_i)$ is the contribution of training data x_i on test instance x_t. The theorem below we call latent representer theorem shows that such decomposition holds for any stationary

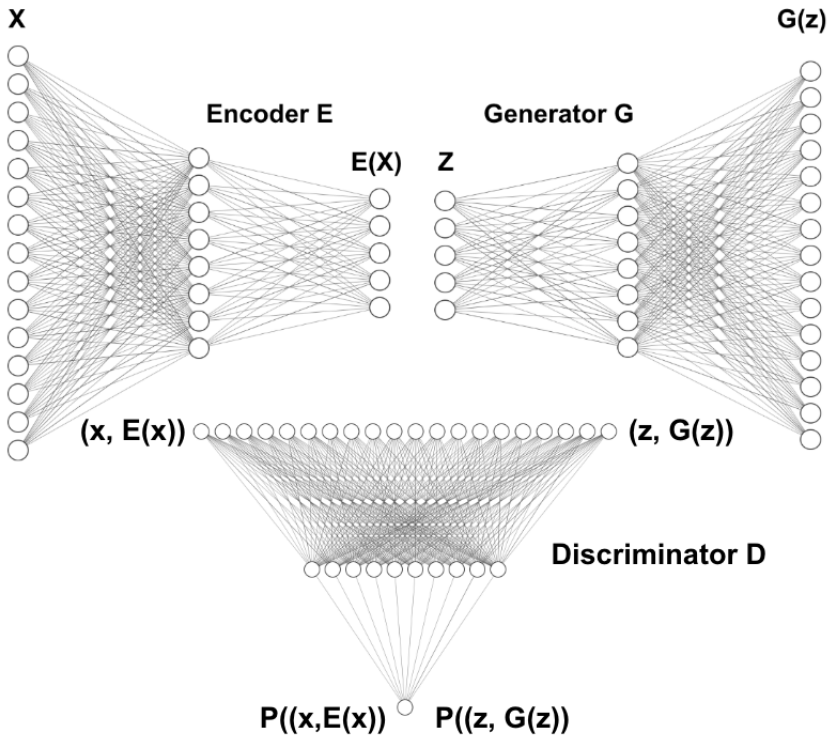


Figure 4.1: Bidirectional GAN visualization.

point solution and any deep neural network. We also address some nice theoretical properties of latent representer theorem in next few theorems.

THEOREM 4.1. *Denote the neural network prediction function by* $\hat{y} = \sigma(\phi(z, \theta))$. $\phi(z, \theta)$ *is a linear function of* f*, i.e., there exists* θ_1 *such that* $\phi(z, \theta) = \theta_1 f$*. The optimization problem is to minimize the loss function with* l_2 *regularization on* θ_1*:*

$$\frac{1}{n}\sum_{i=1}^{n} L(z_i, y_i, \theta) + \lambda\|\theta_1\|_2. \tag{4.1}$$

Denote θ^* *to be a stationary solution to* (4.1)*, then we have the decomposition:*

$$\phi(z_t, \theta^*) = \sum_{i=1}^{n} k(z_i, z_t, \alpha_i), \tag{4.2}$$

where $\alpha_i = \frac{1}{-2\lambda n}\frac{\partial L(z_i, y_i, \theta)}{\partial \phi(z_i, \theta)}$ *and* $k(z_t, z_i, \alpha_i) = \alpha_i f_i^T f_t$*.*

Proof. Since θ^* is a stationary point, then the gradient of (4.1) with repect to θ_1 is 0 when $\theta_1 = \theta^*$. Therefore we have

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial L(z_i, y_i, \theta)}{\partial \theta_1} + 2\lambda\theta_1^* = 0.$$

By chain rule, we have

$$\frac{\partial L(z_i, y_i, \theta)}{\partial \theta_1} = \frac{L(z_i, y_i, \theta)}{\partial \phi(z_i, \theta)}\frac{\partial \phi(z_i, \theta)}{\partial \theta_1} = -2\lambda n \alpha_i f_i^T.$$

Hence we obtain

$$\theta_1^* = -\frac{1}{2\lambda n}\sum_{i=1}^{n}\frac{\partial L(z_i, y_i, \theta)}{\partial \theta_1} = \sum_{i=1}^{n}\alpha_i f_i^T,$$

which finishes the proof considering $\phi(z_t, \theta^*) = \theta_1^* f_t$. □

Note that θ^* is a summation of $\alpha_i f_i^T$, hence α_i can be viewed as the importance of the latent representation z_i on θ^*. For $k(z_i, z_t, \alpha_i)$, a K dimensional value corresponding to K classes, either case can result in a considerable value of $k(z_i, z_t, \alpha_i)$ on class j: α_{ij} has a significant value; $f_i^T f^t$ is relatively large. Therefore $k(\alpha_i, z_i, z_t)$ is a combination of similarity between f_i and f_t and the gradient of individual loss $L(z_i, y_i, \theta^*)$ on $\phi(z_i, \theta^*)$. More comprehensively, when f_i is close to f_t, indicating z_i is close to z_t, and α_{ij} is large and positive, then $\phi(z_t, \theta^*)_j$ gains support for class j from training point x_i . On the other hand, if α_{ij} is large and negative, then x_i tends to resist the model from making prediction of class j. By representer values, we could classify training data into two types of points: those with negative representer values as inhibitory points and those with positive representer values as excitatory instances.

Assumption 5. *x is generated by its latent representation z by some function G, i.e., $x = G(z) + \epsilon$, where $epsilon$ is a white noise with distribution $N(0, \sigma^2)$.*

Assumption 6. *Given a set of representer values* $\{k(z_i, z_t, \alpha_i)\}_{i=1}^n$ *of a neural network, there exists an* $\epsilon_0 > 0$ *such that* $\min_{1\le i,j\le n} |k(z_i, z_t, \alpha_i) - k(z_j, z_t, \alpha_j)| > \epsilon_0$.

THEOREM 4.2. *Under the setting of Theorem 4.1 and Assumption 5-6, given a latent representation* z_t *of a test point* x_t*, the ranking of representer values* $\{k(z_i, z_t, \alpha_i)\}_{i=1}^n$ *obtained from Theorem 4.1 approximates to the true ranking of representer values.*

Proof. The true representer values denoted by $k_1, k_2, \cdots, k_n$ are decided by a neural network with $G(z_i)$, $i = 1, 2, \cdots, n$ as inputs. Under the setting of Theorem 4.1, the neural network prediction function can be denoted by $\hat{y} = \sigma(\phi_0(G(z), \theta_0))$ and $\phi_0(G(z), \theta_0) = \theta_{10} f_0$. Since f_0 is a function of $G(z)$, we have $\phi_0(G(z), \theta_0) = \theta_{10}\tilde{f}_0(z)$ where $\tilde{f}_0 = f_0 \circ G$. Therefore we transform the original neural network with $G(z)$ as input to the one with latent representation z and we denote $\phi_0(G(z), \theta_0)$ by $\tilde{\phi}(z, \theta_0)$. On the other hand, we can always add an additional neural network $G(z)$ before f and refit the model.

For any $\epsilon > 0$, by universal approximation theorem [Hornik et al., 1989], when neural network is complicated enough, we know that there exists a neural network with function $\phi(\cdot, \theta)$ and θ_1 as defined in Theorem 4.1 such that

$$|\tilde{\phi}(z, \theta_0) - \phi(z, \theta)| < \epsilon, \quad \textit{for any } z \in \mathbb{R}^d.$$

Furthermore, we also have

$$|\theta_{10} - C_1\theta_1| = O(\epsilon), \quad |\tilde{f}_0(z) - C_2 f(z)| = O(\epsilon), \tag{4.3}$$

where C_1, C_2 are constant satisfying $C_1 C_2 = 1$.

By (4.3) and $\theta_1 = \sum_{i=1}^n \alpha_i f_i^T$, we have $|\alpha_{i0} - \frac{C_1}{C_2}\alpha_i| = O(\epsilon)$. Therefore for $1 \le i \le n$, it suffices

to show that

$$|k(z_t, z_i, \alpha_i) - \frac{1}{C_2^2} k_i| = O(\epsilon).$$

By Assumption 6, we can always choose ϵ small enough such that $\max_{1 \leq i,j \leq n} |k(z_t, z_i, \alpha_i) - k(z_t, z_j, \alpha_j)| < \epsilon_0$, then the ranking of $\{k(z_t, z_i, \alpha_i)\}_{i=1}^n$ is consistent with that of $\{k_i\}_{i=1}^n$. □

Theorem 4.2 demonstrates the benefit of calibrating representer theorem on latent space, that is, the set of representer values by our method approximates the ranking of true representer values. Since representer values are a decompositin of preactivation value of a test point, the scale is not really important before pushing them through a softmax function. By contrast, the ranking of representer values fairly matters when comparing the impact of training points. The next theorem shows that as long as the encoder E can approximate the true latent representation z very well, the ranking of representer values by our method is still consistent.

Assumption 7. *Suppose the training dataset DX contains n data points and the optimal encoder is denoted by E_n. Then $E_n(x)$ converges to z with probability* 1.

THEOREM 4.3. *Under Assumption 5-7 and the setting of Theorem 4.1, given any encoded latent representation $E_n(x_t)$, the ranking of representer values $\{k(E_n(x_i), E_n(x_t), \alpha_i)\}_1^n$ obtained from Theorem 4.1 converges to the true ranking of representer values in probability.*

Proof. By Assumption 7 and continous mapping theorem, we have

$$\tilde{f}_0(E_n(x)) \xrightarrow{p} \tilde{f}_0(z), \quad f(E_n(x)) \xrightarrow{p} f(z). \tag{4.4}$$

Similary, $\phi_0(E_n(x), \theta) \xrightarrow{p} \phi_0(z, \theta)$. Let θ_n^* be the stationary solution when inputs are $E_n(x)$. We have $\phi_0(z, \theta_n^*) \to \phi_0(z, \theta^*)$, which implies $\theta_n^* \to \theta^*$.

Therefore we have

$$\alpha_{i0}(E_n(x), \theta_n^*)\tilde{f}_{0i}(E_n(x))^T \xrightarrow{p} \alpha_{i0}(z, \theta^*)\tilde{f}_{0i}(z)^T,$$

which finishes the proof considering (4.4) and Theorem 4.2.

□

4.4 Experiments

To demonstrate that our method is able to select insightful representer points, we evaluate on permutation-invariant MNIST LeCun et al.. Scans of human-written digits provide simple but rich features that are easy to understand and analysis, for which reason we think MNIST dataset is an ideal image set to understand and compare recovered representer points. Under permutation invariant setting, each 28×28 digit image is an unstructured 784D vector. We set the latent distribution $p(z) = N(0, 1)$, i.e., a 100D continuous standard normal distribution. We train a bidirectional GAN [Dumoulin et al., 2016] where encoder E and generator G consist of two hidden layers with 512 units and discriminator D consists of three hidden layers with 1024 units. In Figure 4.2, we show sample generations $G(z)$, as well as real data samples x and corresponding BiGAN reconstructions $G(E(x))$. The reconstructions, while certainly imperfect, demonstrate empirically that the BiGAN encoder E and generator G learn approximate inverse mappings. Therefore we have sufficient rationality to take the latent representation $E(x)$ encoded by BiGAN as input in all experiments.

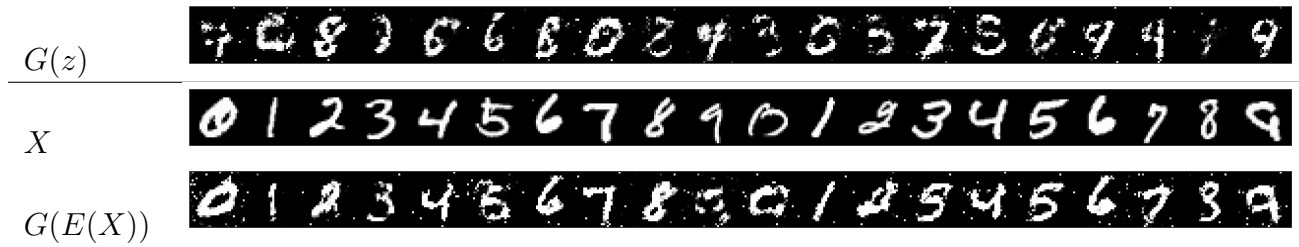


Figure 4.2: BiGAN training results for permutation-invariant MNIST dataset,including generated samples $G(z)$, real data x from digit 0 to 9, and corresponding reconstructions $G(E(x))$.

4.4.1 Dataset Debugging

We repeat the dataset debugging experiment from Yeh et al.. This experiment considers a scenario where humans need to inspect the dataset quality to ensure an improvement of the models performance on the test data. We compare our approach with Yeh et al. on MNIST simulated datasets. In this experiment, we consider handwritten digit "4" and "8" for a binary classification task. We randomly flip 50% of the labels to corrupt the training dataset. Both methods reach a low test accuracy around 0.51. Assume there is a simulated user who checks some fraction of the training data considering some metrics of priority and reflip the wrong labels back. With the partially fixed training data, we retrain the model and compare the test accuracies of each metric.

The L_2 weight decay is set to 10^{-2} for both methods, which is consistent with Yeh et al.. Since test accuracies also slightly depend on randomness of flipping labels, we repeat experiments for 5 random splits for fair comparison. Fraction of flips fixed and corresponding updated test accuracies are summarized in Figure 4.3. We can find that when fraction of training data checked is 5%, 10% and 15%, the simulated user is able to fix slightly more flipped labels by our method. When fraction of flips checked is increasing, both method tend to result in the same amount of flips. Moreover, when fraction of training data checked is small, i.e., 5%, 10%,$\cdots$,40%, our method can achieve much more improvement on the test accuracy than Yeh's method even though it is not evident to tell only from fraction of flips fixed. This implies that our method can select similar amount of wrongly flipped data points but with more "influence" than Yeh's selected ones.

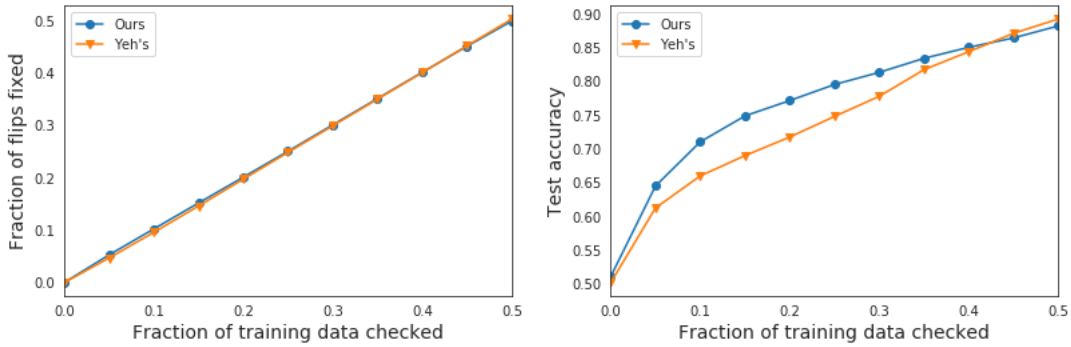


Figure 4.3: Dataset debugging performance of our method and Yeh's method. Our method is able to recover similar amount of flipped training points as Yeh's by inspecting representer value (left) but achieves a far better accuracy (right) after refitting the model with fraction of training data checked as $0.05, 0.10, \cdots, 0.35$.

4.4.2 Image exploring

We are interested in visualizing influential training points with high absolute representer values for correctly classified points and compare the selected top training points with the ones selected from Yeh's method. We still use 10-class MNIST dataset [LeCun et al., 1998] considering that handwritten digit images are simple and easy to recognize and analyze the pattern while other datasets such as CIFAR-10 [Krizhevsky and Hinton, 2009] and Animals with Attributes (AwA) [Lampert et al., 2009] might cause unnecessary challenges. In this experiment, we use a multi-layer perceptrons, where one hidden layer with 512 units and ReLU activation function are adopted. The L_2 weight decay is set to 0.003 for both methods for fair comparison. Both methods achieve test accuracy around 98%.

We pick a few test points both our method and Yeh's models got correct and report top three influential points as shown in Figure 4.4 and 4.5. In Figure 4.4, an image of handwritten digit 3, our method selects three positive images in the same class, and all three images capture the pattern of the test image. Moreover, we notice that the top three images selected by our method ranks by similarity. By Yeh's method, three images in the same class are returned. However, on the one hand, the one ranked 3rd by our method turns to be rank 1st by Yeh's method and the other two

images stay far away from the test image. On the other hand, it is not evident to tell the rationality of the ranking considering similarity. We believe that the one within the same class and is most similar to the test image has most influence on the classification. For the negative examples, in Figure 4.4, the first images picked by both methods are not very similar to the test image but Yeh's method recovers a worse one. The second images by two methods are consistent, implying that both methods can recover some common inhibitory examples.

In Figure 4.5, for positive examples, top 1st images selected by both method are close to the test image and somehow capture the crookedness of digit 2. However, top 2nd and top 3rd images recovered by Yeh's method are not too related to the test image while our method keeps the crookedness. For negative examples, both methods recover similar-looking images with different labels in first two images but different order. However, the 2rd image from Yeh looks fairly unrelated to the test image while ours does. These excitatory and inhibitory examples provide valuable insights for understanding model behavior since they explicitly indicate training points that contribute to the network decision on a particular label for a given test point. Our method can recover training points with more accurate importance compared to Yeh's method in terms of the ranking of similarity. More heuristic examples can be found in the supplementary material.

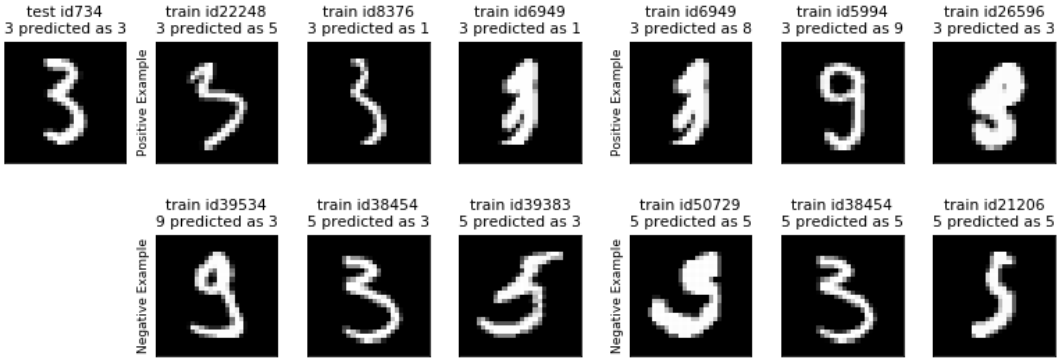


Figure 4.4: Comparison of top three excitatory and inhibitory influential training images for a test point(ID 734) (left-most column) using our method (left columns) and Yeh's method (right columns).

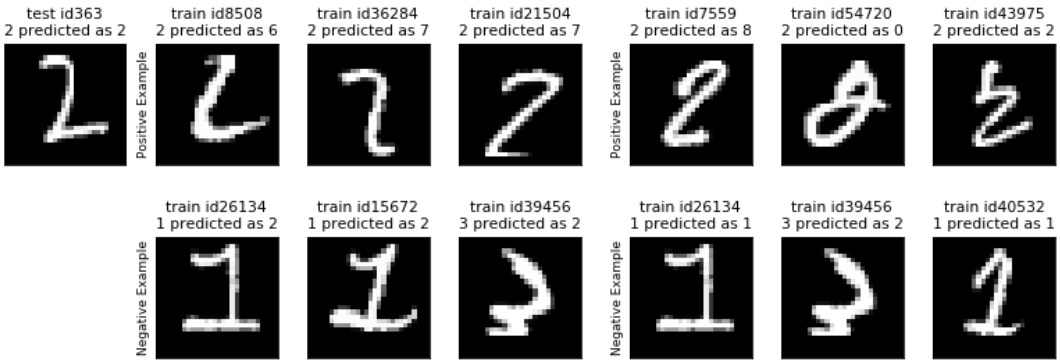


Figure 4.5: Comparison of top three excitatory and inhibitory influential training images for a test point(ID 363) (left-most column) using our method (left columns) and Yeh's method (right columns).

4.4.3 Understanding misclassified image

In this experiment, we select a few test points that are misclassified by our method and Yeh's method and want to understand which training points contribute to fooling the classifier. In the first row of Figure 4.6, consider a handwritten digit 3 predicted as 5, we are interested in two types of training points, one of which encourages the model to make prediction as 5, the other one of which confuses model to inhibit digit 3. These two types of training points are both trying to fool the model but have different emphasis on contribution to the test point. The 2nd and 4th column in 4.6 are training images that most encourage the model to make prediction of 5 by our method and Yeh's method respectively. It is not hard to see that the image recovered by our method looks more similar to the test point, however, it is labeled as 5, which explains why the test point is classified as 5. The 3rd and 5th column give the second type of training images, i.e., the ones resisting classification as 3. Yeh's method returns the same image as ours.

In the second row of 4.6, we have a test image of digit 4 but predicted as 9 by both methods. From 2rd of 4.6, the image selected by our methods fairly similar to the test image but in a different class, which forces the model to make prediction to this different class. However, Yeh's method recovers an image that exactly look like 9. Moreover, the training image resisting prediction of 4

recovered by our method is still close to the test image whereas the one from Yeh's method looks more like digit 1.

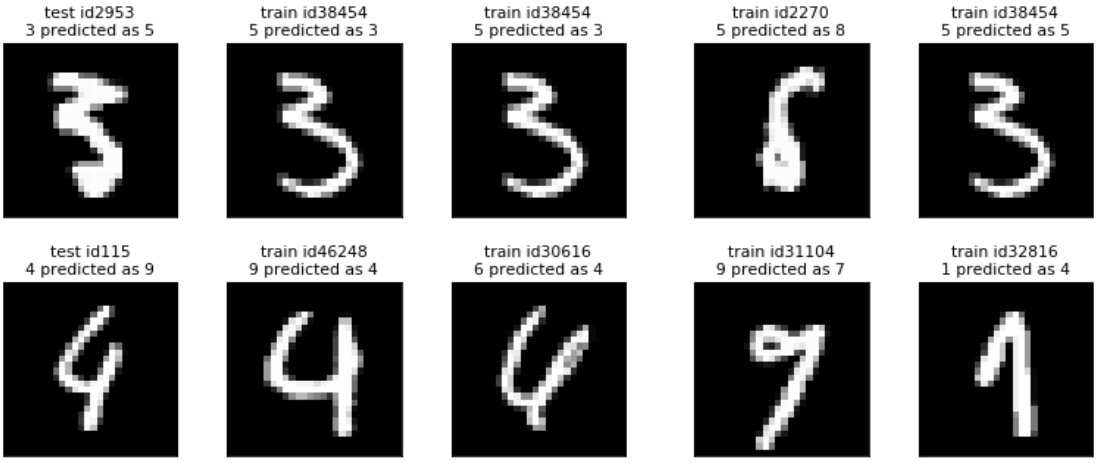


Figure 4.6: A misclassified test image (left most) and the most influential training point by our method(2nd column) and Yeh's method (4th column) supporting the wrongly predicted label and the most influential training point resisting the true label by our method(3rd column) and Yeh's method (5th column).).

4.4.4 Manifold visualization

In this experiment, we are interested in visualizing influential top 3 training images of correctly classified test point on a 2-D scatterplot to further evaluate two methods. The MNIST images have 28×28 pixels, which is hard to be visualized by a 2-D scatterplot. Maaten and Hinton provides t-distributed Stochastic Neighbor Embedding (t-SNE) which has the capacity to characterize much of the local structure revealed in high-dimensional data. Meanwhile, it also allows global structure such as the presence of clusters at several scales. The main idea of t-SNE is to employ a Student-t distribution rather than a Gaussian for computing the similarity between pairwise points, alleviating the optimization and crowding problem existing in Stochastic Neighbor Embedding(SNE) [Hinton and Roweis, 2003].

To speed up the heavy computation caused by calculating pairwise distances between data points, we firstly apply PCA to reduce the dimensionality of the data to 50. Then we use t-SNE on 50-

dimensional representation and show the resulting map as a 2-D scatterplot. Figure 4.7 shows two scatterplot corresponding to two test points as shown in Figure 4.4 and 4.5 respectively. Each training point recovered is annotated with a number indicating its corresponding rank. In the left plot of Figure 4.7, even though our method recovers a training point that is slightly distant from the test point, top 1 point by our method is closest to the test point. In the right plot, the recovered training points by our method locate closer to the test point than the ones selected by Yeh's method. Moreover, the ranking of three recovered points by our method is consistent with the spacial distance whereas Yeh's method fails in keeping the ranking. This is as expected since our method can recover the training images that have much of potential positive contribution to the test image on the latent space.

Figure 4.7: Comparison of t-SNE visualization of recovered training points by our method and Yeh's method for two test point with ID 734(top) and ID 363(bottom).

4.5 Discussion

In this paper we introduced a novel approach of selecting representer training points, those which are influential to a certain model prediction on semantic latent space. We incorporate a modified representer theorem in spirit to representer theorem for empirical risk minimization in RKHS

while calibrating in low-dimensional latent space. This gives rise to a decomposition of a certain prediction value into a sum of representer values contributed from each training point. By the sign and significance of represent value, we can identify a training point to be an excitory or inhibitory point. The optimization has an explicit solution and is sufficiently scalable compared with influence functions as discussed in Yeh et al.. Besides, we also discuss several theoretical properties of our method such as ranking consistency, which supports the rational of the calibration in latent space. Our method has a wide application in diagnosing machine learning such as data debugging. It provides a rich understanding of black-box model behaviors through the view of training data, which is believed to become a standard part of diagnosing machine learning [Koh and Liang, 2017].

In our work, we encode the data into the latent representations as input of the neural network. One potential extension of the work is to find a mapping from representer values in latent space to the ones in arbitrary data space. This would allow more space in model selection, either the target model or encoding model, while keeping the consistency of the ranking of representer values.

Bibliography

Theodore W Anderson. The integral of a symmetric unimodal function over a symmetric convex set and some probability inequalities. *Proceedings of the American Mathematical Society*, 6(2): 170–176, 1955.

Ferran Diego Andilla and Fred A Hamprecht. Sparse space-time deconvolution for calcium image analysis. In *Advances in Neural Information Processing Systems*, pages 64–72, 2014.

Noah Apthorpe, Alexander Riordan, Robert Aguilar, Jan Homann, Yi Gu, David Tank, and H Sebastian Seung. Automatic neuron detection in calcium imaging data using convolutional networks. In *Advances in Neural Information Processing Systems*, pages 3270–3278, 2016.

F. R. Bach. Consistency of trace norm minimization. *Journal of Machine Learning Research*, 8: 1019–1048, 2008a.

Francis R Bach. Consistency of trace norm minimization. *Journal of Machine Learning Research*, 9(Jun):1019–1048, 2008b.

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert MÃžller. How to explain individual classification decisions. *Journal of Machine Learning Research*, 11(Jun):1803–1831, 2010.

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

Michael J Berridge. Neuronal calcium signaling. *Neuron*, 21(1):13–26, 1998.

Bharat Biswal, F Zerrin Yetkin, Victor M Haughton, and James S Hyde. Functional connectivity in the motor cortex of resting human brain using echo-planar MRI. *Magnetic Resonance in Medicine*, 34(4):537–541, 1995.

Bharat B Biswal, Maarten Mennes, Xi-Nian Zuo, Suril Gohel, Clare Kelly, Steve M Smith, Christian F Beckmann, Jonathan S Adelstein, Randy L Buckner, Stan Colcombe, et al. Toward discovery science of human brain function. *Proceedings of the National Academy of Sciences*, 107(10):4734–4739, 2010.

Bastian Bohn, Christian Rieger, and Michael Griebel. A representer theorem for deep kernel learning. *Journal of Machine Learning Research*, 20(64):1–32, 2019.

Mariusz Bojarski, Philip Yeres, Anna Choromanska, Krzysztof Choromanski, Bernhard Firner, Lawrence Jackel, and Urs Muller. Explaining how a deep neural network trained with end-to-end learning steers a car. *arXiv preprint arXiv:1704.07911*, 2017.

Gabriel Cadamuro, Ran Gilad-Bachrach, and Xiaojin Zhu. Debugging machine learning models. In *ICML Workshop on Reliable Machine Learning in the Wild*, 2016.

Jian-Feng Cai, Emmanuel J Candès, and Zuowei Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2010.

Vince D Calhoun and Tulay Adali. Time-varying brain connectivity in fMRI data: whole-brain data-driven approaches for capturing and characterizing dynamic states. *IEEE Signal Processing Magazine*, 33(3):52–66, 2016.

Vince D Calhoun, Robyn Miller, Godfrey Pearlson, and Tulay Adalı. The chronnectome: time-varying connectivity networks as the next frontier in fMRI data discovery. *Neuron*, 84(2):262–274, 2014.

Catie Chang, Zhongming Liu, Michael C Chen, Xiao Liu, and Jeff H Duyn. EEG correlates of time-varying bold functional connectivity. *NeuroImage*, 72:227–236, 2013.

Kun Chen, Hongbo Dong, and Kung-Sik Chan. Reduced rank regression via adaptive nuclear norm penalization. *Biometrika*, 100(4):901–920, 2013a.

Kun Chen, Hongbo Dong, and Kung-Sik Chan. Reduced rank regression via adaptive nuclear norm penalization. *Biometrika*, 100(4):901–920, 2013b.

Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F Stewart, and Jimeng Sun. Doctor ai: Predicting clinical events via recurrent neural networks. In *Machine Learning for Healthcare Conference*, pages 301–318, 2016.

Jan K Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. Attention-based models for speech recognition. In *Advances in neural information processing systems*, pages 577–585, 2015.

Anupam Datta, Shayak Sen, and Yair Zick. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *2016 IEEE symposium on security and privacy (SP)*, pages 598–617. IEEE, 2016.

A. P. Dawid. Some matrix-variate distribution theory: notational considerations and a Bayesian application. *Biometrika*, 68(1):265–274, 1981. ISSN 0006-3444. doi: 10.1093/biomet/68.1.265. URL `http://dx.doi.org/10.1093/biomet/68.1.265`.

Jan Casper De Munck, Hilde M Huizenga, Lourens J Waldorp, and RA Heethaar. Estimating stationary dipoles from MEG/EEG data contaminated with spatially and temporally correlated background noise. *IEEE Transactions on Signal Processing*, 50(7):1565–1572, 2002.

Shanshan Ding and Dennis Cook. Matrix variate regressions and envelope models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(2):387–408, 2018.

Carl Doersch. Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908*, 2016.

Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. Adversarial feature learning. *arXiv preprint arXiv:1605.09782*, 2016.

Richard O Duda, Peter E Hart, and David G Stork. *Pattern Classification*. John Wiley & Sons, 2012.

Vincent Dumoulin, Ishmael Belghazi, Ben Poole, Olivier Mastropietro, Alex Lamb, Martin Arjovsky, and Aaron Courville. Adversarially learned inference. *arXiv preprint arXiv:1606.00704*, 2016.

Bradley Efron. The estimation of prediction error: covariance penalties and cross-validation. *Journal of the American Statistical Association*, 99(467):619–632, 2004.

Amanda Elton and Wei Gao. Task-related modulation of functional connectivity variability and its behavioral correlations. *Human Brain Mapping*, 36(8):3260–3272, 2015.

Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001.

Jianqing Fan, Yang Feng, and Xin Tong. A road to classification in high dimensional space: the regularized optimal affine discriminant. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(4):745–771, 2012.

Michael D Fox, Abraham Z Snyder, Justin L Vincent, Maurizio Corbetta, David C Van Essen, and Marcus E Raichle. The human brain is intrinsically organized into dynamic, anticorrelated functional networks. *Proceedings of the National Academy of Sciences*, 102(27):9673–9678, 2005.

Karl J Friston. Functional and effective connectivity in neuroimaging: a synthesis. *Human Brain Mapping*, 2(1-2):56–78, 1994.

X. Gao, W. Shen, J. Hu, N. Fortin, R. Frostig, and H. Ombao. Regularized matrix data clustering and its application to image analysis. *arXiv:1808.01749*, 2019a.

X. Gao, W. Shen, B. Shahbaba, N. Fortin, and H. Ombao. Evolutionary state-space model and its application to time-frequency analysis of local field potentials. *Statistica Sinica*, 2019b.

Xu Gao, Weining Shen, Chee-Ming Ting, Steven C Cramer, Ramesh Srinivasan, and Hernando Ombao. Modeling brain connectivity with graphical models on frequency domain. *arXiv:1810.03279*, 2018.

Gary H Glover. Overview of functional magnetic resonance imaging. *Neurosurgery Clinics*, 22 (2):133–139, 2011.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

Christine Grienberger and Arthur Konnerth. Imaging calcium in neurons. *Neuron*, 73(5):862–885, 2012.

Shuhang Gu, Lei Zhang, Wangmeng Zuo, and Xiangchu Feng. Weighted nuclear norm minimization with application to image denoising. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2862–2869, 2014.

Frank R. Hampel. *Robust statistics: the approach based on influence functions*, volume 196. Wiley-Interscience, 1986.

Changhee Han, Leonardo Rundo, Ryosuke Araki, Yujiro Furukawa, Giancarlo Mauri, Hideki Nakayama, and Hideaki Hayashi. Infinite brain mr images: Pggan-based data augmentation for tumor detection. *arXiv preprint arXiv:1903.12564*, 2019.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Fritjof Helmchen and Winfried Denk. Deep tissue two-photon microscopy. *Nature Methods*, 2 (12):932, 2005.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, pages 6626–6637, 2017.

Geoffrey E Hinton and Sam T Roweis. Stochastic neighbor embedding. In *Advances in neural information processing systems*, pages 857–864, 2003.

Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

Barry Horwitz. The elusive concept of brain connectivity. *NeuroImage*, 19(2):466–470, 2003.

David T Jones, Prashanthi Vemuri, Matthew C Murphy, Jeffrey L Gunter, Matthew L Senjem, Mary M Machulda, Scott A Przybelski, Brian E Gregg, Kejal Kantarci, David S Knopman, et al. Non-stationarity in the "resting brain's" modular architecture. *PLOS ONE*, 7(6):e39731, 2012.

Gajendra Jung Katuwal and Robert Chen. Machine learning model interpretability for precision medicine. *arXiv preprint arXiv:1610.09045*, 2016.

George S Kimeldorf and Grace Wahba. A correspondence between bayesian estimation on stochastic processes and smoothing by splines. *The Annals of Mathematical Statistics*, 41(2):495–502, 1970.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1885–1894. JMLR. org, 2017.

Dehan Kong, Baiguo An, Jingwen Zhang, and Hongtu Zhu. L2RM: Low-rank linear regression models for high-dimensional matrix responses. *Journal of the American Statistical Association*, page to appear, 2019.

Chayakrit Krittanawong, HongJu Zhang, Zhen Wang, Mehmet Aydar, and Takeshi Kitai. Artificial intelligence in precision cardiovascular medicine. *Journal of the American College of Cardiology*, 69(21):2657–2664, 2017.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

Aaron Kucyi and Karen D Davis. Dynamic functional connectivity of the default mode network tracks daydreaming. *NeuroImage*, 100:471–480, 2014.

Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 951–958. IEEE, 2009.

Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner, et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Nora Leonardi, Jonas Richiardi, Markus Gschwind, Samanta Simioni, Jean-Marie Annoni, Myriam Schluep, Patrik Vuilleumier, and Dimitri Van De Ville. Principal components of functional connectivity: a new approach to study dynamic brain connectivity during rest. *NeuroImage*, 83: 937–950, 2013.

Rongjian Li, Wenlu Zhang, Heung-Il Suk, Li Wang, Jiang Li, Dinggang Shen, and Shuiwang Ji. Deep learning based imaging data completion for improved brain disease diagnosis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 305–312. Springer, 2014.

Raphaël Liégeois, Erik Ziegler, Christophe Phillips, Pierre Geurts, Francisco Gómez, Mohamed Ali Bahri, BT Thomas Yeo, Andrea Soddu, Audrey Vanhaudenhuyse, Steven Laureys, et al. Cerebral functional connectivity periodically (de) synchronizes with anatomical constraints. *Brain Structure and Function*, 221(6):2985–2997, 2016.

Stan Lipovetsky and Michael Conklin. Analysis of regression in game theory approach. *Applied Stochastic Models in Business and Industry*, 17(4):319–330, 2001.

Fabien Lotte, Marco Congedo, Anatole Lécuyer, Fabrice Lamarche, and Bruno Arnaldi. A review of classification algorithms for eeg-based brain–computer interfaces. *Journal of neural engineering*, 4(2):R1, 2007.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, pages 4765–4774, 2017.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

Tara M Madhyastha and Thomas J Grabowski. Age-related differences in the dynamic architecture of intrinsic networks. *Brain Connectivity*, 4(4):231–241, 2014.

Qing Mai, Hui Zou, and Ming Yuan. A direct approach to sparse discriminant analysis in ultra-high dimensions. *Biometrika*, 99(1):29–42, 2012.

S Mika. *Kernel Fisher Discriminant*. PhD thesis, University of Technology, Berlin, 2002.

Y. Mu and F. Gage. Adult hippocampal neurogenesis and its role in alzheimers disease. *Molecular Neurodegeneration*, 6:85, 2011.

S. N. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science*, 27: 538–557, 2012.

Yurii Nesterov. A method of solving a convex programming problem with convergence rate $o(1/k^2)$. In *Soviet Mathematics Doklady*, volume 27, pages 372–376, 1983.

Milad Zafar Nezhad, Dongxiao Zhu, Xiangrui Li, Kai Yang, and Phillip Levy. Safs: A deep feature selection approach for precision medicine. In *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 501–506. IEEE, 2016.

Dong Nie, Roger Trullo, Jun Lian, Caroline Petitjean, Su Ruan, Qian Wang, and Dinggang Shen. Medical image synthesis with context-aware generative adversarial networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 417–425. Springer, 2017.

MM Paul, DH Garry, and TM Edward. Application of fmri in translational medicine and clinical practice journal of nature reviews. *Neuroscience, Nature Publishing Group*, pages 732–744, 2006.

Ashley Petersen, Noah Simon, and Daniela Witten. Scalpel: Extracting neurons from calcium imaging data. *The Annals of Applied Statistics*, 12(4):2430, 2018.

Bettina Platt and Gernot Riedel. The cholinergic system, eeg and sleep. *Behavioural brain research*, 221(2):499–504, 2011.

Maria Giulia Preti, Thomas AW Bolton, and Dimitri Van De Ville. The dynamic functional connectome: State-of-the-art and perspectives. *Neuroimage*, 160:41–54, 2017.

G. Raskutti and M. Yuan. Convex regularization for high-dimensional tensor regression. Technical report, arXiv:1512.01215, 2015.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144. ACM, 2016.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

ME Saab and Jean Gotman. A system to detect the onset of epileptic seizures in scalp eeg. *Clinical Neurophysiology*, 116(2):427–442, 2005.

Ünal Sakoğlu, Godfrey D Pearlson, Kent A Kiehl, Y Michelle Wang, Andrew M Michael, and Vince D Calhoun. A method for evaluating dynamic functional network connectivity and task-modulation: application to schizophrenia. *Magnetic Resonance Materials in Physics, Biology and Medicine*, 23(5-6):351–366, 2010.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in neural information processing systems*, pages 2234–2242, 2016.

Tim Salimans, Han Zhang, Alec Radford, and Dimitris Metaxas. Improving gans using optimal transport. *arXiv preprint arXiv:1803.05573*, 2018.

Bernhard Schölkopf, Ralf Herbrich, and Alex J Smola. A generalized representer theorem. In *International conference on computational learning theory*, pages 416–426. Springer, 2001.

David W Scott. *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons, 2015.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626, 2017.

Jun Shao, Yazhen Wang, Xinwei Deng, and Sijian Wang. Sparse linear discriminant analysis by thresholding for high dimensional data. *The Annals of Statistics*, 39(2):1241–1265, 2011. doi: 10.1214/10-AOS870.

Boris Sharchilev, Yury Ustinovsky, Pavel Serdyukov, and Maarten de Rijke. Finding influential training samples for gradient boosted decision trees. *arXiv preprint arXiv:1802.06640*, 2018.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3145–3153. JMLR. org, 2017.

David Slepian. The one-sided barrier problem for gaussian noise. *Bell System Technical Journal*, 41(2):463–501, 1962.

Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014.

Charles M Stein. Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9(6):1135–1151, 1981.

Enzo Tagliazucchi, Frederic Von Wegner, Astrid Morzelewski, Verena Brodbeck, and Helmut Laufs. Dynamic bold functional connectivity in humans and its electrophysiological correlates. *Frontiers in Human Neuroscience*, 6, 2012.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 58(1):267–288, 1996.

Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108, 2005.

Michael Unser. A representer theorem for deep neural networks. *arXiv preprint arXiv:1802.09210*, 2018.

Aad W Van der Vaart. *Asymptotic Statistics*, volume 3. Cambridge University Press, 2000.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

Xiao Wang and Hongtu Zhu. Generalized scalar-on-image regression models via total variation. *Journal of the American Statistical Association*, 112(519):1156–1168, 2017.

Xiao Wang, Hongtu Zhu, and Alzheimers Disease Neuroimaging Initiative. Generalized scalar-on-image regression models via total variation. *Journal of the American Statistical Association*, 112(519):1156–1168, 2017.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017.

Chih-Kuan Yeh, Joon Kim, Ian En-Hsu Yen, and Pradeep K Ravikumar. Representer point selection for explaining deep neural networks. In *Advances in Neural Information Processing Systems*, pages 9291–9301, 2018.

Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.

T. Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 32(1):56–85, 2004.

X L Zhang, H Begleiter, B Porjesz, W Wang, and A Litke. Event related potentials during object recognition tasks. *Brain Research Bulletin*, 38:531–538, 1995a.

Xiao Lei Zhang, Henri Begleiter, Bernice Porjesz, Wenyu Wang, and Ann Litke. Event related potentials during object recognition tasks. *Brain Research Bulletin*, 38(6):531–538, 1995b.

P. Zhao and B. Yu. On model selection consistency of lasso. *Journal of Machine Learning Research*, 7:2541–2563, 2006.

Wenxuan Zhong and Kenneth S Suslick. Matrix discriminant analysis with application to colorimetric sensor array data. *Technometrics*, 57(4):524–534, 2015.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.

Hua Zhou and Lexin Li. Regularized matrix regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(2):463–483, 2014a.

Hua Zhou and Lexin Li. Regularized matrix regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(2):463–483, 2014b.

Hongtu Zhu, Linglong Kong, Runze Li, Martin Styner, Guido Gerig, Weili Lin, and John H Gilmore. FADTTS: functional analysis of diffusion tensor tract statistics. *NeuroImage*, 56 (3):1412–1425, 2011.

Hongtu Zhu, Runze Li, and Linglong Kong. Multivariate varying coefficient model for functional responses. *Annals of Statistics*, 40(5):2634, 2012.

Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

Appendix A

Appendix for Chapter 2

A.1 Primary lemmas and propositions

We start with some useful lemmas in this section. The proof of main theorems are given in the Appendix B.

We first re-state a singular value thresholding formula in Cai et al. [2010]. This result is extremely useful when computing optimal solution of (A.2), by which the important block of Nestorov's algorithm was formed. The proof is based on showing that 0 is one of subgradients of (A.1) at $\hat{\mathbf{B}}$.

Proposition 3. *For any $\omega \geq 0$ and a given matrix $\mathbf{B}_0 \in \mathcal{R}^{p\times q}$ with singular value decomposition $Udiag(s)V^{\mathrm{T}}$, the minimizer $\hat{\mathbf{B}}$ of*

$$\frac{1}{2}\|\mathbf{B}-\mathbf{B}_0\|_F^2+\omega\|\mathbf{B}\|_* \tag{A.1}$$

has the same singular vectors as B_0 with singular values $(s_i-\omega)_+$.

Next we state a lemma on the risk bound. This result can be viewed as an analog of Theorem 1 in Negahban et al. [2012] under our situation.

Lemma 1. *Suppose that (A1) and (A2) hold, and* $\omega_n \geq 2\|\frac{1}{n}\sum_{i=1}^n \epsilon_i \mathbf{X}_i\|_2$. *Then any optimal solution* $\hat{\mathbf{B}}$ *to*

$$(\hat{\beta}_0, \hat{\mathbf{B}}) = \arg\min_{\beta_0, \mathbf{B}} \frac{1}{2n} \sum_{i=1}^n \Big(y_i - \beta_0 - \langle \mathbf{X}_i, \mathbf{B}\rangle\Big)^2 + \omega_n \|\mathbf{B}\|_* \tag{A.2}$$

satisfies the bound

$$\|\hat{\mathbf{B}} - \mathbf{B}_0\|_F^2 \leq 9\frac{\omega_n^2}{\lambda_l} r.$$

Proof. We apply Theorem 1 in Negahban et al. [2012] to our situation. Observe that the nuclear norm is decomposable, and the squared error loss satisfies $\tau_{\mathcal{L}}(\mathbf{B}_0) = 0$ in that paper. Moreover, the dual norm $\mathcal{R}^*$ to the nuclear norm is simply the spectral norm. The curvature constant $\kappa_{\mathcal{L}}$ in the restricted strong convexity (RSC) condition can be chosen as $\lambda_l^{1/2}$ because the squared error loss is used and the Hessian matrix $\mathrm{E}\{\mathrm{vec}(\mathbf{X})\mathrm{vec}(\mathbf{X})^{\mathrm{T}}\} = \mathbf{\Sigma}_{xx} \geq \lambda_l I$. For a subspace M that contains matrices of the rank at most r, its subspace compatibility constant satisfies

$$\psi(M) = \sup_{\mathbf{U} \in M \setminus \{0\}} \frac{\|\mathbf{U}\|_*}{\|\mathbf{U}\|_F} = \sup_{\mathbf{U} \in M \setminus \{0\}} \frac{\sum_{i=1}^r \sigma_i(\mathbf{U})}{(\sum_{i=1}^r \sigma_i(\mathbf{U})^2)^{1/2}} \leq \sqrt{r},$$

where the last inequality follows by Cauchy-Schwarz inequality. Hence subspace compatibility constant under the low-rank assumption (A2) is bounded by $\sqrt{r}$. □

Next we state a few commonly used lemmas regarding the concentration property and tail probability inequalities of Gaussian and sub-Gaussian random variable (matrices). Their proofs can be found in standard textbooks, e.g., Wainwright [2019].

Lemma 2. *(Hoeffding bound) Suppose that the variables* $\mathbf{X}_i$, $i = 1, 2, \ldots, n$ *are independent and*

X_i has mean μ_i and sub-Gaussian parameter $\mathbf{\Sigma}_i$. Then for all $t \geq 0$, we have

$$P\left(\sum_{i=1}^{n}(\mathbf{X}_i - \mu_i) \geq t\right) \leq \exp(-\frac{t^2}{2\sum_{i=1}^{n}\mathbf{\Sigma}_i^2})$$

Lemma 3. *Assume $\mathbf{X}_1, \ldots, \mathbf{X}_n \in \mathbb{R}^{p\times q}$ are i.i.d. random matrices. Suppose that $\|\mathbf{X}_1\|_2 \leq M$ almost surely, then with probability greater than $1-\delta$,*

$$\left\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i - E\mathbf{X}_1\right\|_2 \leq \frac{6M}{\sqrt{n}}\left(\sqrt{\log\min(p,q)} + \sqrt{\log(1/\delta)}\right)$$

Lemma 4. *Let $\mathbf{A}$ be an $p \times q$ matrix whose entries are independent standard normal random variables. Denote $s_{\min}(\mathbf{A})$ and $s_{\max}(\mathbf{A})$ as smallest singular value and largest singular value of $\mathbf{A}$ respectively. Assume $p \geq q$ without loss of generality. Then*

$$\sqrt{p} - \sqrt{q} \leq E_{s_{min}}(\mathbf{A}) \leq E_{s_{max}}(\mathbf{A}) \leq \sqrt{p} + \sqrt{q}.$$

Lemma 5. *Let $\mathbf{Y} \sim N(0, I_{d\times d})$ be a d-dimensional Gaussian random variable. Then for any function F: $\mathcal{R}^d \to \mathcal{R}$ with Lipschitz constant L, i.e. $|F(\mathbf{x}) - F(\mathbf{y})| \leq L\|\mathbf{x}-\mathbf{y}\|$ for all $\mathbf{x}, \mathbf{y} \in \mathcal{R}^d$, we have*

$$P\left\{|F(\mathbf{Y}) - E(F(\mathbf{Y}))| \geq t\right\} \leq 2\exp(-\frac{t^2}{2L^2}),$$

for any $t > 0$.

Lemma 6. *(Anderson's comparison inequality [Anderson, 1955]) Let $\mathbf{X}$ and $\mathbf{Y}$ be zero-mean Gaussian random vectors with covariance $\mathbf{\Sigma_X}$ and $\mathbf{\Sigma_Y}$ respectively. If $\mathbf{\Sigma_X} - \mathbf{\Sigma_Y}$ is positive semi-definite then for any convex symmetric set C,*

$$P(\mathbf{X} \in C) \leq P(\mathbf{Y} \in C).$$

The following lemma is very useful in establishing rank estimation consistency.

Lemma 7. *Assume (A1) and (A2) hold. Let* $\hat{\mathbf{B}}$ *be a global minimizer of* (A.2)*. If* $n^{1/2}\omega_n$ *tends to* $+\infty$ *and* ω_n *tends to zero, then* $\omega_n^{-1}(\hat{\mathbf{B}} - \mathbf{B}_0)$ *converges in probability to the unique global minimizer* $\boldsymbol{\Delta}$ *of*

$$\min_{\boldsymbol{\Delta}\in \mathrm{R}^{p\times q}} \frac{1}{2} vec(\boldsymbol{\Delta})^{\mathrm{T}} \boldsymbol{\Sigma} vec(\boldsymbol{\Delta}) + tr\{\mathbf{U}_0^{\mathrm{T}} \boldsymbol{\Delta} \mathbf{V}_0\} + \|\mathbf{U}_{0\perp}^{\mathrm{T}} \boldsymbol{\Delta} \mathbf{V}_{0\perp}\|_*.$$

Moreover, $\hat{\mathbf{B}} = \mathbf{B}_0 + \omega_n \boldsymbol{\Delta} + O_p\big(\omega_n \min(p,q) n^{-1/2} + \min(p,q) n^{-1/2} + \omega_n^2 \min(p,q)^{1/2} n^{-1/2}\big)$.

Proof. We can write $\hat{\mathbf{B}} = \mathbf{B}_0 + \omega_n \hat{\boldsymbol{\Delta}}$, where $\hat{\boldsymbol{\Delta}}$ is the global minimum of

$$V_n(\boldsymbol{\Delta}) = \frac{1}{2} \mathrm{vec}(\boldsymbol{\Delta})^{\mathrm{T}} \hat{\boldsymbol{\Sigma}}_{xx} \mathrm{vec}(\boldsymbol{\Delta}) - \omega_n^{-1} \mathrm{tr} \boldsymbol{\Delta}^{\mathrm{T}} \hat{\boldsymbol{\Sigma}}_{\mathbf{X}\epsilon} + \omega_n^{-1}(\|\mathbf{B}_0 + \omega_n \boldsymbol{\Delta}\|_* - \|\mathbf{B}_0\|_*),$$

where $\hat{\boldsymbol{\Sigma}}_{xx} = n^{-1} \sum_{i=1}^n \mathrm{vec}(\mathbf{X}_i)\mathrm{vec}(\mathbf{X}_i)^{\mathrm{T}}$ and $\hat{\boldsymbol{\Sigma}}_{\mathbf{X}\epsilon} = n^{-1} \sum_{i=1}^n \epsilon_i \mathrm{vec}(\mathbf{X}_i)$. Then $\mathrm{vec}(\boldsymbol{\Delta})^{\mathrm{T}} \hat{\boldsymbol{\Sigma}}_{xx} \mathrm{vec}(\boldsymbol{\Delta})/2 - \mathrm{vec}(\boldsymbol{\Delta})^{\mathrm{T}} \boldsymbol{\Sigma}_{xx} \mathrm{vec}(\boldsymbol{\Delta})/2$ converges to $\mathrm{vec}(\boldsymbol{\Delta})^{\mathrm{T}} \mathrm{E}(\hat{\boldsymbol{\Sigma}}_{xx} - \boldsymbol{\Sigma}_{xx}) \mathrm{vec}(\boldsymbol{\Delta})/2$ with probability of 1. Note that $\mathrm{E}\|\hat{\boldsymbol{\Sigma}}_{xx} - \boldsymbol{\Sigma}\|_F^2 = O(n^{-1})$. Denote $\mathrm{vec}(\boldsymbol{\Delta})_i$ as a_i and $(\hat{\boldsymbol{\Sigma}}_{xx} - \boldsymbol{\Sigma})_{ij}$ as b_{ij}. Then we have

$$\begin{aligned}
\frac{1}{2} |\mathrm{vec}(\boldsymbol{\Delta})^{\mathrm{T}} \mathrm{E}(\hat{\boldsymbol{\Sigma}}_{xx} - \boldsymbol{\Sigma}) \mathrm{vec}(\boldsymbol{\Delta})| &\le \sum_{i,j=1}^{pq} |a_i a_j \mathrm{E}(b_{ji})| \\
&\le \Big(\sum_{i,j=1}^{pq} a_i^2 a_j^2 \sum_{i,j=1}^{pq} \mathrm{E}(b_{ij}^2) \Big)^{\frac{1}{2}} \\
&= \sum_{i=1}^{pq} a_i^2 \mathrm{E}(\sum_{i,j=1}^{pq} b_{ij}^2)^{\frac{1}{2}} \\
&= \sum_{i=1}^{pq} a_i^2 \mathrm{E}\|\hat{\boldsymbol{\Sigma}}_{xx} - \boldsymbol{\Sigma}_{xx}\|_F \\
&= \|\boldsymbol{\Delta}\|_F^2 O(n^{-1/2}) \\
&= O\big(\min(p,q) \|\boldsymbol{\Delta}\|_2^2 n^{-1/2} \big).
\end{aligned}$$

Meanwhile,

$$\begin{aligned}|\mathrm{tr}\boldsymbol{\Delta}^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}_{\mathbf{X}\epsilon}| &\leq (\mathrm{tr}\boldsymbol{\Delta}^{\mathrm{T}}\boldsymbol{\Delta})^{\frac{1}{2}}(\mathrm{tr}\hat{\boldsymbol{\Sigma}}_{\mathbf{X}\epsilon}^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}_{\mathbf{X}\epsilon})^{\frac{1}{2}} \\ &= \|\boldsymbol{\Delta}\|_F O_p(n^{-1/2}) \\ &\leq \min(p,q)^{\frac{1}{2}}\|\boldsymbol{\Delta}\|_2 O_p(n^{-\frac{1}{2}}).\end{aligned}$$

Therefore

$$\begin{aligned}V_n(\boldsymbol{\Delta}) =& \frac{1}{2}\mathrm{vec}(\boldsymbol{\Delta})^{\mathrm{T}}\boldsymbol{\Sigma}\mathrm{vec}(\boldsymbol{\Delta}) + O_p\big(\min(p,q)n^{-1/2}\|\boldsymbol{\Delta}\|_2^2\big) + O_p\big(\min(p,q)^{\frac{1}{2}}\omega_n^{-1}n^{-1/2}\|\boldsymbol{\Delta}\|_2\big) \\ &+ \mathrm{tr}(\mathbf{U}_0^{\mathrm{T}}\boldsymbol{\Delta}\mathbf{V}_0) + \|\mathbf{U}_{0\perp}^{\mathrm{T}}\boldsymbol{\Delta}\mathbf{V}_{0\perp}\|_* + O_p\big(\omega_n p^{1/2}q^{1/2}\min(p,q)\|\boldsymbol{\Delta}\|_2^2\big). \\ =& V(\boldsymbol{\Delta}) + O_p\big(\min(p,q)n^{-1/2}\|\boldsymbol{\Delta}\|_2^2\big) + O_p\big(\min(p,q)^{\frac{1}{2}}\omega_n^{-1}n^{-1/2}\|\boldsymbol{\Delta}\|_2\big) \\ &+ O_p\big(\omega_n p^{1/2}q^{1/2}\min(p,q)\|\boldsymbol{\Delta}\|_2^2\big),\end{aligned}$$

where $p^{1/2}q^{1/2}$ in the last term comes from the Frobenius norm of any matrix in $\mathcal{R}^{p\times q}$ with bounded entries. Let s_r be the r-th largest singular value of B_0, for any $M < s_r/(2\omega_n)$,

$$\begin{aligned}&\mathrm{E}\sup_{\|\boldsymbol{\Delta}\|_2\leq M}|V_n(\boldsymbol{\Delta}) - V(\boldsymbol{\Delta})| \\ &= O\big(\min(p,q)M^2\mathrm{E}\|\hat{\boldsymbol{\Sigma}}_{xx} - \boldsymbol{\Sigma}\|_F + M\min(p,q)^{\frac{1}{2}}\omega_n^{-1}E(\|\hat{\boldsymbol{\Sigma}}_{M\epsilon}\|^2)^{1/2} + \omega_n p^{1/2}q^{1/2}\min(p,q)M^2\big) \\ &= fO\big(\min(p,q)M^2n^{-1/2} + M\min(p,q)^{\frac{1}{2}}\omega_n^{-1}n^{-1/2} + \omega_n p^{1/2}q^{1/2}\min(p,q)M^2\big).\end{aligned}$$

Obviously $V(\boldsymbol{\Delta})$ achieves its minimum in the bounded ball at $\boldsymbol{\Delta}_0 \neq 0$. Hence by Markov inequality the probability of the minimum of $V_n(\boldsymbol{\Delta})$ lying strictly inside the ball $\|\boldsymbol{\Delta}\|_2 < 2\|\boldsymbol{\Delta}_0\|_2$ tends to one and is also the unconstrained minimum. □

The following two lemmas can be viewed as analogs of Proposition 3 and Lemma 11 in Bach [2008a]. W present them without the proof.

Lemma 8. *Let* $\mathbf{B}_0 = \mathbf{U}_0 Diag(S_0)\mathbf{V}_0^{\mathrm{T}}$ *be the singular value decomposition of* $\mathbf{B}_0$*. Then the unique*

global minimizer of

$$\frac{1}{2}vec(\mathbf{\Delta})^{\mathrm{T}}\mathbf{\Sigma}vec(\mathbf{\Delta}) + tr\mathbf{U}_0^{\mathrm{T}}\mathbf{\Delta}\mathbf{V}_0 + \|\mathbf{U}_{0\perp}^{\mathrm{T}}\mathbf{\Delta}\mathbf{V}_{0\perp}\|_*$$

satisfies $\mathbf{U}_{0\perp}^{\mathrm{T}}\mathbf{\Delta}\mathbf{V}_{0\perp} = 0$ *if and only if*

$$\left\|\{(\mathbf{V}_{0\perp}\otimes\mathbf{U}_{0\perp})^{\mathrm{T}}\mathbf{\Sigma}^{-1}(\mathbf{V}_{0\perp}\otimes\mathbf{U}_{0\perp})\}^{-1}\{(\mathbf{V}_{0\perp}\otimes\mathbf{U}_{0\perp})^{\mathrm{T}}\mathbf{\Sigma}^{-1}(\mathbf{V}_0\otimes\mathbf{U}_0)vec(\mathbf{I})\}\right\|_2 \leq 1.$$

Furthermore, when $\mathbf{U}_{0\perp}^{\mathrm{T}}\mathbf{\Delta}\mathbf{V}_{0\perp} = 0$*, the solutions has these forms:*

$$vec(\Lambda) = \{(\mathbf{V}_{0\perp}\otimes\mathbf{U}_{0\perp})^{\mathrm{T}}\mathbf{\Sigma}(\mathbf{V}_{0\perp}\otimes\mathbf{U}_{0\perp})\}^{-1}\{(\mathbf{V}_{0\perp}\otimes\mathbf{U}_{0\perp})^{\mathrm{T}}\mathbf{\Sigma}(\mathbf{V}_0\otimes\mathbf{U}_0)vec(I)\},$$

$$vec(\mathbf{\Delta}) = -\mathbf{\Sigma}^{-1}vec(\mathbf{U}_0\mathbf{V}_0^{\mathrm{T}} - \mathbf{U}_{0\perp}\Lambda\mathbf{V}_{0\perp}^{\mathrm{T}}). \tag{A.3}$$

Lemma 9. *The matrix* $\mathbf{B}$ *with singular value decomposition* $\mathbf{B} = \mathbf{U}Diag(\mathbf{S})\mathbf{V}^{\mathrm{T}}$*(with strictly positive singular value s) is optimal for the problem in* (A.2) *if and only if*

$$\hat{\mathbf{\Sigma}}_{xx}\mathbf{B} - \hat{\mathbf{\Sigma}}_{\mathbf{X}y} + \omega_n\mathbf{U}\mathbf{V}^{\mathrm{T}} + N = 0,$$

with $\mathbf{U}^{\mathrm{T}}\mathbf{N} = 0$*,* $\mathbf{N}\mathbf{V} = 0$ *and* $\|\mathbf{N}\|_2 \leq \omega_n$*.*

A.2 Proof of Theorems

Proof of Theorem 2.1. Throughout the proof, we use C to denote a universal positive constant where its value is not important for the theoretical purpose. In order to apply Lemma 1, we just need to evaluate the term $\|n^{-1}\sum_{i=1}^n \epsilon_i\mathbf{X}_i\|_2$ and then set the tuning parameter w_n to be greater than

that quantity. Note that $\epsilon_i = Y_i - \langle \mathbf{X}_i, \mathbf{B} \rangle - \beta_0^*$. Let $\mathbf{X}_i = \pi_1 \mathbf{X}_i^{(1)} + \pi_2 \mathbf{X}_i^{(2)}$, where $\text{vec}(\mathbf{X}_i^{(g)}) \overset{i.i.d.}{\sim} N(\mu_g, \mathbf{\Sigma})$ and $\mu_g \in \mathbb{R}^{pq \times 1}$ for $g = 1, 2$. Define $\mathbf{X}_i \overset{\text{i.i.d}}{\sim} \mathbf{X}$ and $\epsilon_i \overset{\text{i.i.d}}{\sim} \epsilon$. Observe that

$$
\begin{aligned}
&\text{vec}\{\text{E}(\epsilon_i \mathbf{X}_i)\} \\
&= \pi_1 \text{E}\{(-\frac{n}{n_1} - \beta_0^* - \langle \mathbf{X}^{(1)}, \mathbf{B}_0 \rangle)\text{vec}(\mathbf{X}^{(1)})\} + \pi_2 \text{E}\{(\frac{n}{n_2} - \beta_0^* - \langle \mathbf{X}^{(2)}, \mathbf{B}_0 \rangle)\text{vec}(\mathbf{X}^{(2)})\} \\
&= (\mu_2 - \mu_1) - (\pi_1 \mu_1 + \pi_2 \mu_2)\beta_0^* - \pi_1 E\{\text{vec}(\mathbf{X}^{(1)})\text{vec}(\mathbf{X}^{(1)})^{\text{T}}\}\text{vec}(\mathbf{B}_0) \\
&\qquad - \pi_2 E\{\text{vec}(\mathbf{X}^{(2)})\text{vec}(\mathbf{X}^{(2)})^{\text{T}}\}\text{vec}(\mathbf{B}_0) \\
&= (\mu_2 - \mu_1) - (\pi_1 \mu_1 + \pi_2 \mu_2)\beta_0^* - \pi_1 \{\mu_1 \mu_1^{\text{T}} + \mathbf{\Sigma}\}\text{vec}(\mathbf{B}_0) - \pi_2 \{\mu_2 \mu_2^{\text{T}} + \mathbf{\Sigma}\}\text{vec}(\mathbf{B}_0).
\end{aligned}
\tag{A.4}
$$

Now, to further simplify this result, we reparameterize the mean of two normal populations such that $\mu_1 = 0$, and $\mu_2 = \mathbf{D}$. Then recall by the equivalence between LDA and least squares solution, we have

$$
\text{vec}(\mathbf{B}) = c\mathbf{\Sigma}^{-1}\mathbf{D}, \;\; \beta_0 = -(\pi_1 \mu_1 + \pi_2 \mu_2)^{\text{T}}\text{vec}(\mathbf{B}) = -\pi_2 c \mathbf{D}^{\text{T}} \mathbf{\Sigma}^{-1} \mathbf{D}, \;\; \beta_0^* = \beta_0 - d
$$

for some positive constants c and d. Then (A.4) can be simplified into

$$
\begin{aligned}
&\mathbf{D} - \pi_2 \mathbf{D}\beta_0^* - \pi_2 \{\mathbf{D}\mathbf{D}^{\text{T}}\}\text{vec}(\mathbf{B}) - c\mathbf{D} \\
&= \mathbf{D} - \pi_2 \mathbf{D}\beta_0 + \pi_2 \mathbf{D}\mathbf{D} - \pi_2 \{\mathbf{D}\mathbf{D}^{\text{T}}\}\text{vec}(\mathbf{B}) - c\mathbf{D} \\
&= \mathbf{D}\{1 + \pi_2^2 c \mathbf{D}^{\text{T}} \mathbf{\Sigma}^{-1} \mathbf{D} + \pi_2 d - \pi_2 c \mathbf{D}^{\text{T}} \mathbf{\Sigma}^{-1} \mathbf{D} - c\} \\
&= 0,
\end{aligned}
$$

given d is chosen as $\pi_2^{-1}\{c - 1 + (\pi_2 - \pi_2^2)(\mathbf{D}^{\text{T}} \mathbf{\Sigma}^{-1} \mathbf{D})\}$.

Next we show that with high probability, $\|\epsilon \mathbf{X}\|_2 \leq 2 \log n (C_\mu + \lambda_u^{1/2})(\sqrt{p} + \sqrt{q} + \sqrt{\log n})$. Since ϵ follows a mixture of two normal distributions, ϵ is sub-gaussian with sub-gaussian parameter denoted by σ, which is a positive constant due to the bounded eigenvalue condition in (A1). By

Lemma 2, for sufficiently large n,

$$P(|\epsilon| > 2\log n) \leq P(|\epsilon - E(\epsilon)| > \log n) \leq 2\exp(-\frac{\log^2 n}{2\sigma^2}) \leq C\exp(-2\log n) = \frac{C}{n^2}.$$

Then we know $|\epsilon| \leq 2\log n$ with probability of at least $1 - Cn^{-2}$. For $\|\mathbf{X}\|_2$, we first consider its centralized version, that is, $\mathbf{X} \sim N(0, \mathbf{\Sigma})$. Note that we can write the spectral norm of a matrix in the form of a canonical Gaussian process,

$$\|N(0, \mathbf{\Sigma})\|_2 = \sup_{\mathbf{A}:\|\mathbf{A}\|_* \leq 1} \langle N(0, \mathbf{\Sigma}), \mathbf{A}\rangle.$$

This allows us to apply Gaussian comparison inequality [Slepian, 1962]. Define $\mathbf{Z} \in \mathbb{R}^{p\times q}$ that satisfies $\text{vec}(\mathbf{Z}) \sim N(0, \mathbf{I})$. Then by Lemma 6, we have

$$\begin{aligned} P(\|N(0, \mathbf{\Sigma})\|_2 > t_1) &= P\Big(\sup_{A:\|A\|_* \leq 1} \langle N(0, \mathbf{\Sigma}), \mathbf{A}\rangle > t_1\Big) \\ &\leq P\Big(\sup_{\mathbf{A}:\|\mathbf{A}\|_* \leq 1} \langle \mathbf{Z}, \mathbf{A}\rangle > t_1\lambda_u^{-1/2}\Big) \\ &= P(\|\mathbf{Z}\|_2 > t_1\lambda_u^{-1/2}) \end{aligned} \tag{A.5}$$

for any $t_1 > 0$ because $\mathbf{\Sigma} \leq \lambda_u\mathbf{I}$ due to (A1). Apply Lemma 5 (or more generally the Tracy-Widow law), we have

$$P(\|\mathbf{Z}\|_2 - E\|\mathbf{Z}\|_2 > \sqrt{\log n}) \leq C\exp(-2\log n) = Cn^{-2}$$

for some constant $C > 0$. Since $E\|\mathbf{Z}\|_2 \leq \sqrt{p} + \sqrt{q}$, by Lemma 4, with probability of at least $1 - Cn^{-2}$, $\|\mathbf{Z}\|_2 \leq \sqrt{p} + \sqrt{q} + \sqrt{\log n}$, which leads to $\|N(0, \mathbf{\Sigma})\|_2 \leq \lambda_u^{1/2}(\sqrt{p} + \sqrt{q} + \sqrt{\log n})$

by (A.5). Therefore with probability of at least $1 - Cn^{-2}$,

$$\begin{aligned}
\|\epsilon\mathbf{X}\|_2 &\leq (2\log n)\|\mathbf{X}\|_2 \\
&\leq 2\log n(\|\mu_1\|_2 + \|N(0,\mathbf{\Sigma})\|_2) \\
&\leq 2\log n\left\{C_\mu(\sqrt{p}+\sqrt{q}) + \lambda_u^{1/2}(\sqrt{p}+\sqrt{q}+\sqrt{\log n})\right\} \\
&\leq 2\log n(C_\mu + \lambda_u^{1/2})(\sqrt{p}+\sqrt{q}+\sqrt{\log n})
\end{aligned}$$

using Condition (A4) and since we assume $\mu_2 = 0$ without loss of generality.

Now we apply the standard matrix concentration inequality, (e.g., Lemma 3) with $M = 2\log n(C_\mu + \lambda_u^{1/2})(\sqrt{p} + \sqrt{q} + \sqrt{\log n})$ and $\delta = n^{-1}$. Note that $P(\|\mathbf{X}_i\epsilon_i\|_2 \leq M, \quad i = 1,\ldots,n) = (1 - Cn^{-2})^n \geq 1 - Cn^{-1}$ by Bernoulli's inequality. Hence we obtain that with probability of at least $1 - Cn^{-1}$,

$$\begin{aligned}
\left\|\frac{1}{n}\sum_{i=1}^n \mathbf{X}_i\epsilon_i - \mathrm{E}(\epsilon\mathbf{X})\right\|_2 &\leq \frac{6M}{\sqrt{n}}\left(\sqrt{\log\min(p,q)} + \sqrt{\log 1/\delta}\right) \\
&\leq \frac{12(\log n)^{3/2}(C_\mu + \lambda_u^{1/2})(\sqrt{p}+\sqrt{q}+\sqrt{\log n})}{\sqrt{n}}
\end{aligned}$$

This completes the proof. □

Proof of Theorem 2.2. By Lemma 7, we obtain $\hat{\mathbf{B}} = \mathbf{B}_0 + \omega_n\mathbf{\Delta} + o_p(\omega_n)$. Since the rank is a lower semi-continuous function, the rank of $\hat{\mathbf{B}}$ is larger than r with probability tending to one by the consistency result, where r is the rank of $\mathbf{B}_0$. Suppose $\hat{\mathbf{B}}$ has singular value decomposition USV^{T} and U_c, V_c are singular vectors corresponding to U and V except the r largest singular values. By Lemma 9, $\hat{\mathbf{\Sigma}}_{xx}(\hat{\mathbf{B}} - \mathbf{B}_0) - \hat{\mathbf{\Sigma}}_{\mathbf{X}\epsilon}$ and $\hat{\mathbf{B}}$ have simultaneous singular value decomposition. Therefore it suffices to show $\|\mathbf{U}_c^{\mathrm{T}}\{\hat{\mathbf{\Sigma}}_{xx}(\hat{\mathbf{B}} - \mathbf{B}_0) - \hat{\mathbf{\Sigma}}_{\mathbf{X}\epsilon}\}\mathbf{V}_c\|_2 < \omega_n$ with probability tending to one. Note

that

$$\begin{aligned}\mathbf{U}_c^{\mathrm{T}}\{\hat{\mathbf{\Sigma}}_{xx}(\hat{\mathbf{B}}-\mathbf{B}_0)-\hat{\mathbf{\Sigma}}_{\mathbf{X}\epsilon}\}\mathbf{V}_c &= \mathbf{U}_c^{\mathrm{T}}\{\omega_n\hat{\mathbf{\Sigma}}_{xx}\mathbf{\Delta}+o_p(\omega_n)-O_p(n^{-1/2})\}\mathbf{V}_c \\ &= \omega_n\mathbf{U}_c^{\mathrm{T}}(\mathbf{\Sigma}\mathbf{\Delta})\mathbf{V}_c+o_p(\omega_n),\end{aligned}$$

where $\mathbf{\Sigma}\mathbf{\Delta}$ is the matrix in $\mathrm{R}^{p\times q}$ satisfying $\mathrm{vec}(\mathbf{\Sigma}\mathbf{\Delta})=\mathbf{\Sigma}\mathrm{vec}(\mathbf{\Delta})$. Because of the regular consistency and a positive eigengap for $\mathbf{B}_0$, the projection onto the first singular vectors of $\hat{\mathbf{B}}$ converges those of $\mathbf{B}_0$. Hence the projection on the orthogonal space is also consistent, which means $\mathbf{U}_c\mathbf{U}_c^{\mathrm{T}}$converges to $\mathbf{U}_{0\perp}\mathbf{U}_{0\perp}^{\mathrm{T}}$ and $\mathbf{V}_c\mathbf{V}_c^{\mathrm{T}}$converges to $\mathbf{V}_{0\perp}\mathbf{V}_{0\perp}^{\mathrm{T}}$. Then by Lemma 8, we have

$$\begin{aligned}\|\mathbf{U}_c^{\mathrm{T}}\{\hat{\mathbf{\Sigma}}_{xx}(\hat{\mathbf{B}}-\mathbf{B}_0)-\hat{\mathbf{\Sigma}}_{\mathbf{X}\epsilon}\}\mathbf{V}_c\|_2 &= \|\mathbf{U}_c\mathbf{U}_c^{\mathrm{T}}\{\hat{\mathbf{\Sigma}}_{xx}(\hat{\mathbf{B}}-\mathbf{B}_0)-\hat{\mathbf{\Sigma}}_{\mathbf{X}\epsilon}\}\mathbf{V}_c\mathbf{V}_c^{\mathrm{T}}\|_2 \\ &= \omega_n\|\mathbf{U}_{0\perp}\mathbf{U}_{0\perp}^{\mathrm{T}}(\mathbf{\Sigma}\mathbf{\Delta})\mathbf{V}_{0\perp}\mathbf{V}_{0\perp}^{\mathrm{T}}\|_2+o_p(\omega_n) \\ &= \omega_n\|\mathbf{U}_{0\perp}\mathbf{U}_{0\perp}^{\mathrm{T}}\mathbf{\Sigma}\{-\mathbf{\Sigma}^{-1}(\mathbf{U}_0\mathbf{V}_0{}^{\mathrm{T}}-\mathbf{U}_{0\perp}\Lambda\mathbf{V}_{0\perp}^{\mathrm{T}})\}\mathbf{V}_{0\perp}\mathbf{V}_{0\perp}^{\mathrm{T}}\|_2+o_p(\omega_n) \\ &= \omega_n\|\mathbf{U}_{0\perp}\mathbf{\Lambda}\mathbf{V}_{0\perp}^{\mathrm{T}}\|_2+o_p(\omega_n) \\ &= \omega_n\|\mathbf{\Lambda}\|_2+o_p(\omega_n),\end{aligned}$$

where the third equality is due to (A.3). Since $\|\mathbf{\Lambda}\|_2<1$, the last expression is less than ω_n with probability tending to one, which completes the proof. □

Proof of Theorem 2.3. Based on Corollary 3.1 of Zhang [2004], we have

$$R(\hat{f}_n)\le R^*+2c(\epsilon_1+\epsilon_2)^{1/s},$$

where Q is the squared error loss function defined by $Q(f)=E_{\mathbf{X}}\{y-f(\mathbf{X})^2\}$, $\epsilon_1=\inf_f E_{\mathbf{X}}(2P(Y=1\mid\mathbf{X})-1-f(\mathbf{X}))^2$, ϵ_2 satisfies $Q(\hat{f}_n)\le\inf_f Q(f)+\epsilon_2$, and c and s can be chosen as $c=0.5$ and $s=2$ as explained by the Example 3.1 (for least squares loss function) in that paper. Now note that since $\hat{f}_n$ is determined by the classification coefficient $\hat{\mathbf{B}}$ and $\hat{\beta}_0$ that are both consistent based on Theorem 2.1. Therefore ϵ_2 can be chosen arbitrarily close to 0. Also, as we assume the true class label Y given $\mathbf{X}$ is determined by the linear classification rule with β_0^* and $\mathbf{B}_0$, then

$\inf_f E_{\mathbf{X}}\{2P(Y=1 \mid \mathbf{X}) - 1 - f(\mathbf{X})\}^2 = 0$. Therefore $\epsilon_1 = 0$. This concludes the proof. $\square$

Appendix B

Appendix for Chapter 3

B.1 Useful lemmas

We first re-state some lemmas that will be useful in the proof of the risk bound and rank consistency results. For simplicity of notation, we define

$$L(Y; x) = \frac{1}{2n} \sum_{i=1}^{n} K_H(x - X_i) \|Y_i - Y\|_F^2.$$

Note that sometimes we may write $L(Y; x)$ as $L(Y)$ for simplicity. Let $\nabla L(Y)$ be the gradient of $L(Y)$. Then the solution to (3.4) can be written as

$$\hat{g}(x) = \underset{Y \in \mathbb{R}^{p \times q}}{\arg\min} \{L(Y; x) + \lambda_n R(Y)\},$$

where $R(\cdot)$ is a norm on $\mathbb{R}^{p \times q}$, and we denote $R^*(\cdot)$ as the dual norm of $R(\cdot)$. In our setting, $R(\cdot)$ is the nuclear norm and $R^*(\cdot)$ is the spectral norm.

The first lemma is taken from Negahban et al. [2012]. It provides useful general risk bound for

high-dimensional regularized M-estimators.

Lemma 10. *When $\lambda_n \geq 2R^*(\nabla L(Y))$, any optimal solution $\hat{Y}_{\lambda_n}$ satisfies the bound*

$$\|\hat{Y}_{\lambda_n} - Y\|^2 \leq 9\frac{\lambda_n^2}{k_L^2}\Phi^2(\overline{M}) + \frac{\lambda_n}{k_L}\{2\tau_L^2(Y) + 4R(Y_{M^\perp})\} \tag{B.1}$$

Under our setting, $\Phi(\overline{M}) = \sup_{u \in M\setminus\{0\}} \frac{R(u)}{\|u\|} = r$, $\tau_L(Y) = 0$, $R(Y_{M^\perp}) = 0$, and k_L is a constant related to the strong convexity of the loss function that we will specify in Proposition 3.

The next few lemmas are standard concentration bounds results for sub-Gaussian random variables and Gaussian matrices.

Lemma 11. *Let X be zero-mean, and supported on some interval [a, b] almost surely. Then X is sub-Gaussian with parameter at most $\sigma = b - a$.*

Lemma 12. *(Hoeffding bound) Suppose that $X_1, \ldots, X_n$ are independent and X_i has mean μ_i and sub-Gaussian parameter σ_i for $i = 1, \ldots, n$. Then for every $t \geq 0$, we have*

$$P\left(\sum_{i=1}^n (X_i - \mu_i) \geq t\right) \leq \exp(-\frac{t^2}{2\sum_{i=1}^n \sigma_i^2}).$$

Lemma 13. *(Gordon's theorem for Gaussian matrices) Let A be an $p \times q$ matrix whose entries are independent standard normal random variables. Denote $s_{\min}(A)$ and $s_{\max}(A)$ as the smallest and the largest singular value of A respectively. Assume $p \geq q$ without loss of generality. Then*

$$\sqrt{p} - \sqrt{q} \leq E\{s_{min}(A)\} \leq E\{s_{max}(A)\} \leq \sqrt{p} + \sqrt{q}.$$

Lemma 14. *Let $Y \sim N(0, I_{d\times d})$ be a d-dimensional Gaussian random variable. Then for any function F: $\mathbb{R}^d \to \mathbb{R}$ with Lipschitz constant L, i.e. $|F(x) - F(y)| \leq L\|x - y\|$ for all $x, y \in \mathbb{R}^d$,*

we have for any $t > 0$*,*

$$P\left\{|F(Y) - E(F(Y))| \geq t\right\} \leq 2\exp(-\frac{t^2}{2L^2}).$$

The next few results are taken from Bach [2008b]. They will be used for proving rank consistency.

Lemma 15. *(Subdifferential) Suppose* $Y = UDiag(\sigma)V^T$ *is the singular value decomposition of* Y*, where* $U \in \mathbb{R}^{p\times r}$ *and* $V \in \mathbb{R}^{q\times r}$ *have orthogonal columns and* $\sigma \in \mathbb{R}^r$ *with each element positive, then the subdifferential of* $\|\cdot\|_*$ *is equal to*

$$\|\cdot\|_*(Y) = \{UV^T + M, \text{ such that } \|M\|_2 \leq 1, U^T M = 0 \text{ and } MV = 0\}.$$

Lemma 16. *(Directional derivative) The directional derivative at* $Y = UDiag(\sigma)V^T$ *is equal to:*

$$\lim_{t\to 0^+} \frac{\|Y + t\Delta\|_* - \|Y\|_*}{t} = tr(U^T\Delta V) + \|U_\perp^T \Delta V_\perp\|_*.$$

Lemma 17. *Assume Y has rank* $r < \min\{p, q\}$ *with ordered singular value decomposition* $Y = UDiag(\sigma)V^T$*. If* $\frac{4}{s_r}\|\Delta\|_2^2 < \|(I - UU^T)\Delta(I - VV^T)\|_2$*, then rank* $(Y + \Delta) > r$*.*

Lemma 18. *Assume* Σ *is any invertible matrix and Y has singular value decomposition* $Y = UDiag(\sigma)V^T$*. Then the unique global minimizer of*

$$vec(\Delta)^T\Sigma vec(\Delta) + trU^T\Delta V + \|U_\perp^T\Delta V_\perp\|_*$$

satisfies $U_\perp^T\Delta V_\perp = 0$ *if and only if*

$$vec(\Lambda) = \|((V_\perp \otimes U_\perp)^T\Sigma^{-1}(V_\perp \otimes U_\perp))^{-1}((V_\perp \otimes U_\perp)^T\Sigma^{-1}(V \otimes Uvec(I))\|_2 \leq 1.$$

Moreover, the optimal solution Δ satisfies

$$vec(\Delta) = -\Sigma^{-1} vec(UV^T - U_\perp \Lambda V_\perp^T).$$

B.2 Curvature and strong convexity

One of the major ingredients in the proof for the risk bound result is the strong convexity of the loss function. This is described using $\delta L(Y) = L(Y+\Delta) - L(Y) - < \nabla L, \Delta >$, the remainder of the first-order Taylor expansion along some direction Δ. Then we have the following proposition stating that with high probability, the loss function L is strongly convex.

Proposition 4. *With probability of at least* $1 - \exp(-\frac{nf(x)^2h^{2s}}{32k_{\max}^2})$, $\delta L(Y) \geq k_L \|\Delta\|_F^2$, *where* $k_L = \frac{f(x)}{4} - \frac{C_k}{2} h^{\alpha_1}$.

Proof. To satisfy restricted strong convexity condition, we must have $\delta L(Y) \geq k_L \|\Delta\|_F^2$ for $\Delta \in \zeta(\Delta)$, where $k_L > 0$ is some constant, and $\zeta(\Delta) = \{\Delta | \|\pi_{M^\perp}(\Delta)\|_* \leq 3\|\pi_{\overline{M}}(\Delta)\|_*\}$, where $M^\perp$ and $\overline{M}$ are defined in Section 2.2 of Negahban et al. [2012]. Then we have

$$\begin{aligned}
\delta L(Y) &= L(Y+\Delta) - L(Y) - < \nabla L, \Delta > \\
&= \frac{1}{2n} \sum_{i=1}^n K_H(x - X_i) \|Y_i - Y - \Delta\|_F^2 - \frac{1}{2n} \sum_{i=1}^n K_H(x - X_i) \|Y_i - Y\|_F^2 \\
&\quad + \frac{1}{n} \sum_{i=1}^n K_H(x - X_i) < Y_i - Y - \Delta, \Delta > \\
&= \frac{1}{2n} \sum_{i=1}^n K_H(x - X_i) < \Delta, \Delta >,
\end{aligned}$$

where the inner product $< \cdot, \cdot >$ is defined by $< A, B >= \sum_{i,j=1}^{p,q} A_{ij} B_{ij}$. Therefore we just need to show $\frac{1}{2n} \sum_{i=1}^n K_H(x - X_i) \geq k_L$ with high probability.

By Assumption 3, $K_H(x - X_i) \in [0, h^{-s} k_{\max}]$ for any x and X_i. Then $K_H(x - X_i) - E(K_H(x -$

$X_i)) \in [-h^{-s}k_{\max}, h^{-s}k_{\max}]$.

By Lemma 11, we have

$$E_{X_i}(e^{\lambda(K_H(x-X_i)-E(K_H(x-X_i)))}) \leq e^{\frac{\lambda^2(k_{\max}-(-k_{\max}))^2}{2h^{2s}}} = e^{\frac{4\lambda^2 k_{\max}^2}{2h^{2s}}}.$$

Hence $K_H(x - X_i) - E(K_H(x - X_i))$ is sub-gaussian by definition. By Lemma 12, we further have, for any $t > 0$,

$$P\left\{\frac{1}{n}\sum_{i=1}^{n} K_H(x - X_i) - E(K_H(x - X_i)) \leq -t\right\} \leq \exp\left(-\frac{nt^2h^{2s}}{8k_{\max}^2}\right).$$

Meanwhile,

$$P\left\{\frac{1}{n}\sum_{i=1}^{n} K_H(x - X_i) - (E(K_H(x - X_i)) - f(x)) - f(x) \leq -t\right\} \leq \exp(-\frac{nt^2h^{2s}}{8k_{\max}^2}). \quad \text{(B.2)}$$

By classical kernel density estimation theory, e.g., [Van der Vaart, 2000], there exists $C_k \geq 0$ such that $|E(K_H(x - X_i)) - f(x)| \leq C_k h^{\alpha_1}$. For simplicity, set $t = \frac{f(x)}{2}$ and $k_L = \frac{f(x)}{4} - \frac{C_k}{2}h^{\alpha_1}$, which is positive given h is small enough and $f(x)$ is lower bounded due to Assumption 3. Therefore, with probability of at least $1 - \exp(-\frac{nf(x)^2h^{2s}}{32k_{\max}^2})$, $\frac{1}{2n}\sum_{i=1}^{n} K_H(x - X_i) \geq k_L$. This completes the proof. □

B.3 Proof of Theorem 3.1

Denote $R^*(\cdot)$ as the dual norm of $R(\cdot)$. In our case, $R^*(\cdot)$ is the spectral norm, which is defined as the largest singular value of a matrix. We have

$$\begin{aligned}
R^*(\nabla L(Y,x)) =& R^*\left(\frac{1}{n}\sum_{i=1}^n K_H(x-X_i)(Y_i-Y)\right)\\
=& R^*\left(\frac{1}{n}\sum_{i=1}^n K_H(x-X_i)(g(X_i)+\epsilon_i-g(x)\right)\\
\leq& R^*\left(\frac{1}{n}\sum_{i=1}^n K_H(x-X_i)(g(X_i)-g(x))\right)+R^*\left(\frac{1}{n}\sum_{i=1}^n K_H(x-X_i)\epsilon_i\right).
\end{aligned}$$

Therefore, for any $t>0$,

$$\begin{aligned}
&P\left\{\left\|\frac{1}{n}\sum_{i=1}^n K_H(x-X_i)(Y_i-Y)\right\|_2>t\right\}\\
\leq& P\left\{\left\|\frac{1}{n}\sum_{i=1}^n K_H(x-X_i)\epsilon_i\right\|_2+\left\|\frac{1}{n}\sum_{i=1}^n K_H(x-X_i)(g(X_i)-g(x))\right\|_2>t\right\}\\
\leq& P\left\{\left\|\frac{1}{n}\sum_{i=1}^n K_H(x-X_i)\epsilon_i\right\|_2>t/2\right\}+P\left\{\left\|\frac{1}{n}\sum_{i=1}^n K_H(x-X_i)(g(X_i)-g(x))\right\|_2>t/2\right\}.
\end{aligned}\tag{B.3}$$

First we look at $P\left\{\|\frac{1}{n}\sum_{i=1}^n K_H(x-X_i)\epsilon_i\|_2>t_1\right\}$, where $t_1>0$. Note that

$$E\left(\left\|\frac{1}{n}\sum_{i=1}^n K_H(x-X_i)\epsilon_i\right\|_2\right)=E\left(E\left(\left\|\frac{1}{n}\sum_{i=1}^n K_H(x-X_i)\epsilon_i\right\|_2\mid\{X_i\}_{i=1}^n\right)\right).$$

Denote $K_H(x - X_i)$ as c_i. Then

$$E(\|\frac{1}{n}\sum c_i\epsilon_i\|_2) = E(\frac{\|\sum c_i\epsilon_i\|_2/n}{\sqrt{\sigma^2\sum_{i=1}^n c_i^2/n^2}}) \times \sqrt{\frac{\sigma^2\sum_{i=1}^n c_i^2}{n^2}}$$
$$\leq \sigma\frac{\sqrt{p}+\sqrt{q}}{n} \times (\sum_{i=1}^n c_i^2)^{1/2}, \tag{B.4}$$

where the last inequality is due to Lemma 13 since entries of $\frac{\|\sum_{i=1}^n c_i\epsilon_i\|_2/n}{\sqrt{\sigma^2\sum_{i=1}^n c_i^2/n^2}}$ are i.i.d. standard normal random variables. Then

$$\begin{aligned}
E\left(\|\frac{1}{n}\sum_{i=1}^n K_H(x-X_i)\epsilon_i\|_2\right) &\leq \sigma\frac{\sqrt{p}+\sqrt{q}}{n}E\Big((\sum_{i=1}^n K_H(x-X_i)^2)^{1/2}\Big) \\
&\leq \sigma\frac{\sqrt{p}+\sqrt{q}}{n}\Big(E(\sum_{i=1}^n K_H(x-X_i)^2)\Big)^{1/2} \\
&= \sigma\frac{\sqrt{p}+\sqrt{q}}{\sqrt{n}}\Big(EK_H(x-X_1)^2\Big)^{1/2} \\
&\leq 2\sigma\frac{\sqrt{p}+\sqrt{q}}{\sqrt{n}}\Big(EK_H(x-X_1)^2 I(\|x-X_i\|_\infty = O(h))\Big)^{1/2} \\
&\leq 2\sigma\frac{\sqrt{p}+\sqrt{q}}{\sqrt{n}}\Big(\frac{k_{\max}^2}{h^{2s}}P(\|x-X_i\|_\infty = O(h))\Big)^{1/2} \\
&\leq 2\sigma\frac{\sqrt{p}+\sqrt{q}}{\sqrt{n}}\Big(\frac{k_{\max}^2}{h^{2s}}C_f h^s\Big)^{1/2} \\
&= 2\sigma k_{\max}C_f^{1/2}\frac{\sqrt{p}+\sqrt{q}}{\sqrt{n}h^{s/2}},
\end{aligned}$$

where we have used the fact that $K_H(x - X_i)$ is negligible once $\|x - X_i\|_\infty \gg h$ and $P\Big(\|x -$

$X_i\|_\infty = O(h)\Big) = C_f h^s$ since the density function of x is bounded from above. Then we have

$$
\begin{aligned}
&P\left\{\|\frac{1}{n}\sum_{i=1}^{n} K_H(x-X_i)\epsilon_i\|_2 - E(\|\frac{1}{n}\sum_{i=1}^{n} K_H(x-X_i)\epsilon_i\|_2) > t_1\right\} \\
=&E\left(P\left\{\|\frac{1}{n}\sum_{i=1}^{n} K_H(x-X_i)\epsilon_i\|_2 - E(\|\frac{1}{n}\sum_{i=1}^{n} K_H(x-X_i)\epsilon_i\|_2) > t_1 \mid \{X_i\}_{i=1}^n\right\}\right) \\
=&E\Big(P\Big\{\frac{\|\frac{1}{n}\sum_{i=1}^{n} K_H(x-X_i)\epsilon_i\|_2}{(\sigma^2\sum_{i=1}^{n} c_i^2)^{1/2}/n} - E\left(\frac{\|\frac{1}{n}\sum_{i=1}^{n} K_H(x-X_i)\epsilon_i\|_2}{(\sigma^2\sum_{i=1}^{n} c_i^2)^{1/2}/n}\right) \\
&\qquad > \frac{t_1}{(\sigma^2\sum_{i=1}^{n} c_i^2)^{1/2}/n} \mid \{X_i\}_{i=1}^n\Big\}\Big) \\
\leq& E\left(\exp\Big(-\frac{t_1^2}{2\sigma^2\sum_{i=1}^{n} K_H(x-X_i)^2/n^2}\Big)\right) \qquad \text{(B.5)} \\
\leq& \exp\left(-\frac{t_1^2}{4\sigma^2 E\Big(K_H(x-X_i)^2\Big)/n}\right) \qquad \text{(B.6)} \\
=& \exp\left(-\frac{nt_1^2}{4\sigma^2 E\Big(K_H(x-X_i)^2 I(\|x-X_i\|_\infty = O(h))\Big)}\right) \\
\leq& \exp\left(-\frac{nt_1^2 h^s}{4C_f k_{\max}^2\sigma^2}\right). \qquad \text{(B.7)}
\end{aligned}
$$

In the above derivation, (B.5) is due to Lemma 14 and the fact that spectral norm is 1-Lipschitz. Moreover, (B.6) holds since $\frac{\sum_{i=1}^{n} K_H(x-X_i)^2}{n}$ converges to $E(K_H(x-X_i)^2)$ by strong law of large numbers and hence is bounded by $2E(K_H(x-X_i)^2)$ with probability tending to 1 as the variance of $K_H(x-X_1)^2$ is finite.

Then we take a look at the second term in (B.3). By strong law of large numbers, continuity of $\|\cdot\|$

and continuous mapping theorem, we have

$$
\begin{aligned}
&\|\frac{1}{n}\sum_{i=1}^{n} K_H(x-X_i)(g(X_i)-g(x))\|_2 \\
&\le 2\|E\Big(K_H(x-X_i)(g(X_i)-g(x))\Big)\|_2 \\
&\le \|2E\Big(K_H(x-X_i)(g(X_i)-g(x))\Big)\|_F \\
&= \|2E\Big(K_H(x-X_i)(g(X_i)-g(x))I(\|X_i-x\|_\infty = O(h))\Big)\|_F \\
&\le 2\frac{k_{\max}}{h^s}\|E\Big(|g(X_i)-g(x)|I(\|X_i-x\|_\infty = O(h))\Big)\|_F \\
&\le 2\frac{k_{\max}}{h^s}P\{\|X_i-x\|_\infty = O(h)\}\sqrt{pq}CMh^{\alpha_2} \\
&\le 2C_fCMk_{\max}\sqrt{pq}h^{\alpha_2},
\end{aligned}
$$

where M comes from the fact that $\|X_i-x\|_\infty \le Mh$ with probability tending to 1 as $\|X_i-x\|_\infty = O_p(h)$.

By matching $2CC_fMk_{\max}\sqrt{pq}h^{\alpha_2}$ with $2\sigma k_{\max}C_f^{1/2}\frac{\sqrt{p}+\sqrt{q}}{\sqrt{n}h^{s/2}}$, we set $t_1 = 2C_fCMk_{\max}\sqrt{pq}h^{\alpha_2}$. Then

$$
\begin{aligned}
&P\left\{\|\frac{1}{n}\sum_{i=1}^{n} K_H(x-X_i)\epsilon_i\|_2 > 2CC_fMk_{\max}\sqrt{pq}h^{\alpha_2} + \sigma k_{\max}C_f^{1/2}\frac{\sqrt{p}+\sqrt{q}}{\sqrt{n}h^{s/2}}\right\} \\
\le& P\left\{\|\frac{1}{n}\sum_{i=1}^{n} K_H(x-X_i)\epsilon_i\|_2 - E(\|\frac{1}{n}\sum_{i=1}^{n} K_H(x-X_i)\epsilon_i\|_2) > 2CC_fMk_{\max}\sqrt{pq}h^{\alpha_2}\right\} \\
\le& \exp(-\frac{nt_1^2h^s}{4C_fk_{\max}^2\sigma^2}) \\
=& \exp(-\frac{C_fC^2M^2pqnh^{2\alpha_2+s}}{\sigma^2}).
\end{aligned}
$$

Denote $2\sigma k_{\max} C_f^{1/2} \frac{\sqrt{p}+\sqrt{q}}{\sqrt{n}h^{s/2}}$ by t_2. We obtain

$$P\left\{\|\frac{1}{n}\sum_{i=1}^{n} K_H(x-X_i)(Y_i-Y)\|_2 > 2(t_1+t_2)\right\} \leq \exp(-\frac{C_f C^2 M^2 pqnh^{2\alpha_2+s}}{\sigma^2}).$$

By Lemma 10, when $\lambda_n = 4(t_1+t_2)$, we have

$$\|\hat{Y}_{\lambda_n} - Y\|_F^2 \leq 9\frac{16(t_1+t_2)^2}{(\frac{f(x)}{4} - \frac{C_k}{2}h^{\alpha_1})^2} r \leq C\lambda_n^2 r,$$

with probability of at least $1-\exp(-\frac{C_f C^2 M^2 pqnh^{2\alpha_2+s}}{\sigma^2})-\exp(-\frac{nf(x)^2h^{2s}}{32k_{\max}^2})$, where $t_1 = 2C_f CMk_{\max}\sqrt{pq}h^{\alpha_2}$ and $t_2 = 2\sigma k_{\max} C_f^{1/2}\frac{\sqrt{p}+\sqrt{q}}{\sqrt{n}h^{s/2}}$.

B.4 Rank consistency: proof of Theorem 3.2

We first state and prove two useful propositions.

Proposition 5. *$-\frac{1}{n}\sum_{i=1}^{n} K_H(x-X_i)(Y_i-Y)$ and Y have simultaneous singular value decompositions.*

Proof. By Lemma 15, the minimizer of of $L(Y;x)+\lambda R(Y)$ satisfies

$$-\frac{1}{n}\sum_{i=1}^{n} K_H(x-X_i)(Y_i-Y) + \lambda(UV^T+M) = 0.$$

where Y has singular value decomposition $Y = U\text{Diag}(\sigma)V^T$(with strictly positive singular value vector σ) and $U^T N = 0$, $MV = 0$, $\|M\|_2 \leq 1$, which completes the proof. □

Proposition 6. *Let $\hat{g}(x)$ be a global minimizer of Eq.* (3.4) *and assume $\hat{g}(x) = g(x) + \lambda_n\hat{\Delta}$. Then*

$\hat{\Delta}$ converges in probability to the unique global minimizer Δ of

$$\min_{\Delta\in\mathbb{R}^{p\times q}} \frac{f(x)}{2}\|\Delta\|_F^2 + tr(U^T\Delta V) + \|U_\perp^T\Delta V_\perp\|_*.$$

Moreover, we have

$$\begin{aligned}\hat{g}(x) =& g(x) + \lambda_n\Delta + O_p\big(n^{-1/2}h^{-s/2}\lambda_n\min(p,q)\big) + O_p\big(\lambda_n^2\min(p,q)^2\big)I \\ &+ O_p(\sqrt{pq\min(p,q)})(n^{-1/2}h^{-s/2} + h^{\alpha_2}).\end{aligned}$$

Proof. Under the assumption $\hat{g}(x) = g(x) + \lambda_n\hat{\Delta}$, we have that

$$\arg\min_{Y\in\mathbb{R}^{p\times q}}\left\{\frac{1}{2n}\sum_{i=1}^n K_H(x-X_i)\|Y_i - Y\|_F^2 + \lambda_n\|Y\|_*\right\}$$

is equivalent with

$$\begin{aligned}&\arg\min_{\Delta\in\mathbb{R}^{p\times q}}\left\{\frac{1}{2n}\sum_{i=1}^n K_H(x-X_i)\|Y_i - g(x) - \lambda_n\Delta\|_F^2 + \lambda_n\|g(x)+\lambda_n\Delta\|_* - \lambda_n\|(g(x)\|_*\right\} \\ =&\arg\min_{\Delta\in\mathbb{R}^{p\times q}}\Big\{\frac{\lambda_n^2}{2n}\sum_{i=1}^n K_H(x-X_i)\|\Delta\|_F^2 - \frac{\lambda_n}{n}\sum_{i=1}^n K_H(x-X_i) < Y_i - g(x), \Delta > \\ &\qquad + \lambda_n\|g(x) + \lambda_n(\Delta\|_* - \|g(x)\|_*)\Big\} \\ =&\arg\min_{\Delta\in\mathbb{R}^{p\times q}}\Big\{\frac{1}{2n}\sum_{i=1}^n K_H(x-X_i)\|\Delta\|_F^2 - \frac{1}{n\lambda_n}\sum_{i=1}^n K_H(x-X_i) < Y_i - g(x), \Delta > \\ &\qquad + \frac{\|g(x)+\lambda_n\Delta\|_* - \|g(x)\|_*}{\lambda_n}\Big\}.\end{aligned}$$

Denote

$$\begin{aligned}V_n(\Delta) = &\frac{1}{2n}\sum_{i=1}^n K_H(x-X_i)\|\Delta\|_F^2 - \frac{1}{n\lambda_n}\sum_{i=1}^n K_H(x-X_i) < Y_i - g(x), \Delta > \\ &+ \frac{\|g(x)+\lambda_n\Delta\|_* - \|g(x)\|_*}{\lambda_n}.\end{aligned}$$

First we treat the term $\frac{1}{n\lambda_n}\sum_{i=1}^n K_H(x-X_i) < Y_i - g(x), \Delta >$. Note that

$$
\begin{aligned}
&|\frac{1}{n\lambda_n}\sum_{i=1}^n K_H(x-X_i) < Y_i - g(x), \Delta > | \\
\leq &\frac{1}{\lambda_n}\left\|\frac{1}{n}\sum_{i=1}^n K_H(x-X_i)(Y_i-g(x))\right\|_F \|\Delta\|_F \\
\leq &\frac{1}{\lambda_n}\left\|\frac{1}{n}\sum_{i=1}^n K_H(x-X_i)(g(x_i)-g(x))\right\|_F \|\Delta\|_F + \frac{1}{\lambda_n}\|\frac{1}{n}\sum_{i=1}^n K_H(x-X_i)\epsilon_i\|_F\|\Delta\|_F \\
\leq &\frac{2}{\lambda_n}\|E\big(K_H(x-X_i)(g(x_1)-g(x)\big)\|_F\|\Delta\|_F + \frac{1}{\lambda_n}O_p(\frac{\sqrt{pq}}{\sqrt{nh^s}})\|\Delta\|_F \\
\leq &\frac{2C_fCMk_{\max}\sqrt{pq}h^{\alpha_2}}{\lambda_n}\|\Delta\|_F + \frac{1}{\lambda_n}O_p(\frac{\sqrt{pq}}{\sqrt{nh^s}})\|\Delta\|_F.
\end{aligned}
$$

Assume $g(x)$ has singular value decomposition $g(x) = U\text{Diag}(\sigma)V^T$ with positive singular value vector σ, then by Lemma 16,

$$
\frac{\|g(x)+\lambda_n\Delta\|_* - \|g(x)\|_*}{\lambda_n} = \text{tr}(U^T\Delta V) + \|U_\perp^T\Delta V_\perp\|_* + O_p(\lambda_n\|\Delta\|_F^2).
$$

Therefore

$$
\begin{aligned}
V_n(\Delta) =&\frac{f(x)}{2}\|\Delta\|_F^2 + O_p(n^{-1/2}h^{-s/2})\|\Delta\|_F^2 + \text{tr}(U^T\Delta V) + \|U_\perp^T\Delta V_\perp\|_* + O_p(\lambda_n\|\Delta\|_F^2) \\
&+ 2C_fCMk_{\max}\sqrt{pq}h^{\alpha_2}\lambda_n^{-1}|\Delta\|_F + \frac{1}{\lambda_n}O_p(\frac{\sqrt{pq}}{\sqrt{nh^s}})\|\Delta\|_F \\
=&V(\Delta) + O_p(n^{-1/2}h^{-s/2})\|\Delta\|_F^2 + O_p(\lambda_n\|\Delta\|_F^2) + 2C_fCMk_{\max}\sqrt{pq}h^{\alpha_2}\lambda_n^{-1}\|\Delta\|_F \\
&+ \frac{1}{\lambda_n}O_p(\frac{\sqrt{pq}}{\sqrt{nh^s}})\|\Delta\|_F \\
\leq &V(\Delta) + O_p\big(n^{-1/2}h^{-s/2}\min(p,q)\big)\|\Delta\|_2^2 + O_p(\lambda_n\min(p,q)^2\|\Delta\|_2^2) \\
&+ 2C_fCMk_{\max}\sqrt{pq}h^{\alpha_2}\min(p,q)^{1/2}\lambda_n^{-1}\|\Delta\|_2 + \frac{1}{\lambda_n}O_p\left(\frac{\sqrt{pq\min(p,q)}}{\sqrt{nh^s}}\right)\|\Delta\|_2,
\end{aligned}
$$

where $V(\Delta) = \frac{f(x)}{2}\|\Delta\|_F^2 + \text{tr}(U^T\Delta V) + \|U_\perp^T\Delta V_\perp\|_*$. Then we have for any $M_0 > 0$,

$$E \sup_{\|\Delta\|_2 \le M_0} |V_n(\Delta) - V(\Delta)| = O_p\big(n^{-1/2}h^{-s/2}\min(p,q)\big)M_0^2 + O_p(\lambda_n \min(p,q)^2)M_0^2 + \frac{2C_f C M k_{\max}\sqrt{pq}h^{\alpha_2}\min(p,q)^{1/2}}{\lambda_n}M_0 + \frac{1}{\lambda_n}O_p\left(\frac{\sqrt{pq\min(p,q)}}{\sqrt{nh^s}}\right)M_0.$$

Suppose that $V(\Delta)$ reaches the minimum at a bounded point $\Delta_0 \neq 0$. Then by Markov inequality, in the ball $\|\Delta\|_2 \le 2\|\Delta_0\|_2$, $V_n(\Delta)$ reaches its local minimum with probability tending to 1. Since V_n is convex, the local minimum is also a global one. This completes the proof. □

Proof of Theorem 3.2 Let $\hat{g}(x)$ be a global minimizer of Eq.(3.4). In Lemma 18, we can choose Σ as $f(x)/2$, then

$$\|((V_\perp \otimes U_\perp)^T f(x)^{-1}(V_\perp \otimes U_\perp))^{-1}((V_\perp \otimes U_\perp)^T f(x)^{-1}(V \otimes U\text{vec}(I))\|_2 = 0.$$

Therefore the solution of $\min V(\Delta)$ satisfies $U_\perp^T\Delta V_\perp = 0$. Moreover, $\Delta = -2f(x)^{-1}(UV^T)$.

From previous discussion, we have $\hat{g}(x) = g(x) + \lambda_n\Delta + o_p(\lambda_n)$, where $\hat{g}(x)$ has singular value decomposition $\tilde{U}\text{Diag}(\tilde{s})\tilde{V}^T$. We denote $\tilde{U}_0$ and $\tilde{V}_0$ as the singular vectors corresponding to all but the r largest singular values.

Then we have

$$
\begin{aligned}
&- \tilde{U}_0^T \frac{1}{n} \sum_{i=1}^{n} K_H(x - X_i)(Y_i - \hat{g}(x))\tilde{V}_0 \\
=& - \tilde{U}_0^T \Big(\frac{1}{n} \sum_{i=1}^{n} K_H(x - X_i)(g(X_i) - g(x)) + \frac{1}{n} \sum_{i=1}^{n} K_H(x - X_i)\epsilon_i \\
&\quad - \frac{\lambda_n}{n} \sum_{i=1}^{n} K_H(x - X_i)\Delta - \frac{o_p(\lambda_n)}{n} \sum_{i=1}^{n} K_H(x - X_i) \Big) \tilde{V}_0 \\
=& - \tilde{U}_0^T \Big(Ch^{\alpha_2} + O_p(n^{-1/2}h^{-s/2}) - \lambda_n f(x)\Delta - \lambda_n O_p(n^{-1/2}h^{-s/2})\Delta \\
&\quad - o_p(\lambda_n)O_p(n^{-1/2}h^{-s/2}) \Big) \tilde{V}_0 \\
=& \, \tilde{U}_0^T(\lambda_n f(x)\Delta)\tilde{V}_0 + o_p(\lambda_n/\sqrt{pq}),
\end{aligned}
$$

where the last equation is due to assumption that $\sqrt{pq}h^{\alpha_2} \min(p,q)^{1/2}\lambda_n^{-1} \to 0$ and $\lambda_n^{-1} O_p\left(\frac{\sqrt{pq\min(p,q)}}{\sqrt{nh^s}}\right) \to 0$, further indicating $h^{\alpha_2} = o(\lambda_n/\sqrt{pq})$ and $O_p(n^{-1/2}h^{-s/2}) = o_p(1/\sqrt{pq})$ if $p \wedge q$ is finite, and $h^{\alpha_2} = o(\lambda_n/\sqrt{pq(p \wedge q)})$ and $O_p(n^{-1/2}h^{-s/2}) = o_p(1/\sqrt{pq(p \wedge q)})$ if $p \wedge q \to \infty$.

Since $\tilde{U}_0\tilde{U}_0^T$ and $\tilde{V}_0\tilde{V}_0^T$ converge to $U_0U_0^T$ and $V_0V_0^T$ respectively in probability, we have

$$
\begin{aligned}
\left\| \tilde{U}_0^T \frac{1}{n} \sum_{i=1}^{n} K_H(x - X_i)(Y_i - \hat{g}(x))\tilde{V}_0 \right\|_2 &= \left\| \tilde{U}_0\tilde{U}_0^T \frac{1}{n} \sum_{i=1}^{n} K_H(x - X_i)(Y_i - \hat{g}(x))\tilde{V}_0\tilde{V}_0^T \right\|_2 \\
&= \lambda_n f(x) \|U_\perp U_\perp^T \Delta V_\perp V_\perp^T\|_2 + o_p(\lambda_n) \\
&= o_p(\lambda_n).
\end{aligned}
$$

Therefore $\left\| \tilde{U}_0^T \frac{1}{n} \sum_{i=1}^{n} K_H(x - X_i)(Y_i - \hat{g}(x))\tilde{V}_0 \right\|_2$ is strictly less than λ_n with probability tending to one, which means $\text{rank}(\hat{g}(x)) \leq r$. Therefore we obtain rank consistency.

Appendix C

Appendix for Chapter 4

C.1 Supplementary examples of MNIST

In this section, we provide more examples for the experiments of image exploring and understanding misclassified images.

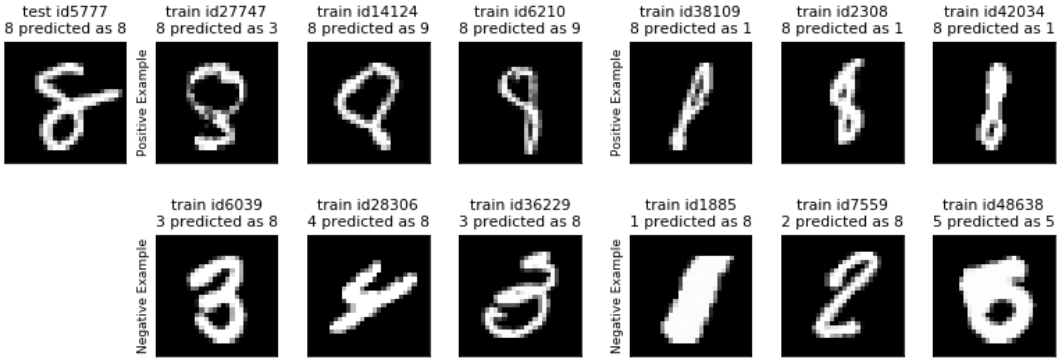


Figure C.1: Comparison of top three excitatory and inhibitory influential training images for a test point(ID 5777) (left-most column) using our method (left columns) and Yeh's method (right columns).

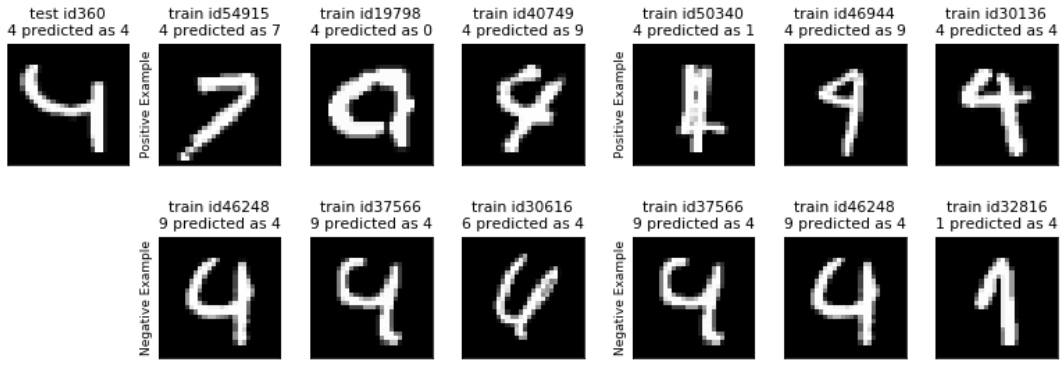


Figure C.2: Comparison of top three excitatory and inhibitory influential training images for a test point(ID 360) (left-most column) using our method (left columns) and Yeh's method (right columns).

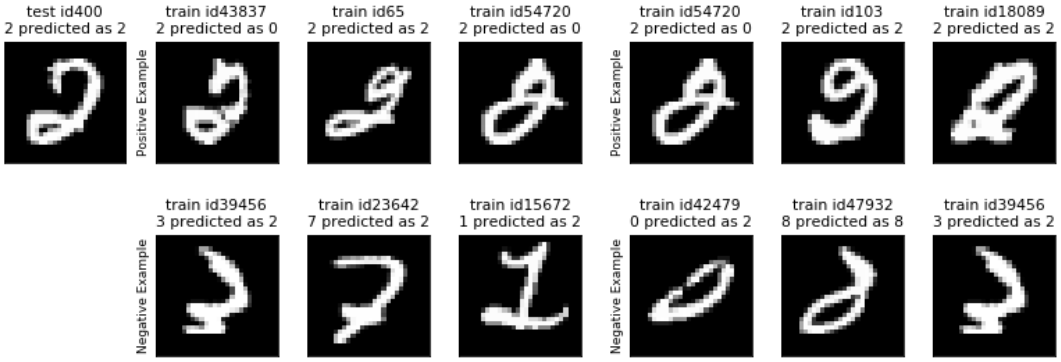


Figure C.3: Comparison of top three excitatory and inhibitory influential training images for a test point(ID 400) (left-most column) using our method (left columns) and Yeh's method (right columns).

C.2 Experiments on Fashion-MNIST dataset

In this section, we apply our method on Fashion-MNIST dataset [Xiao et al., 2017] and repeat a few heuristic experiments as we did on MNIST dataset. Fashion-MNIST consists of 60000 training samples and 10000 test examples. Each sample is processed as a 28×28 grey-scale image from 10 categories of fashion products: (0) T-shirt/top, (1) Trouser, (2) Pullover, (3) Dress, (4) Coat, (5) Sandal (6) Shirt, (7) Sneaker, (8) Bag, (9) Ankle boot. Fashion-MNIST is considered to be more challenging as a drop-in alternative of MINIST for the purpose of classification.

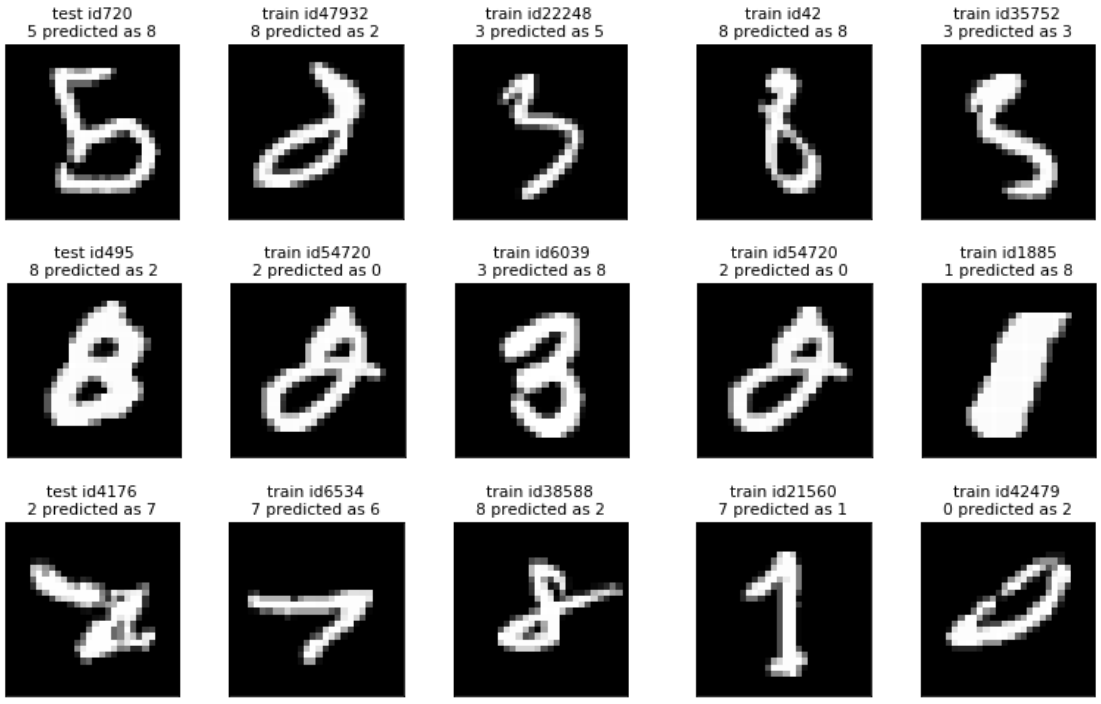


Figure C.4: A misclassified test image (left most) and the most influential training point by our method(2nd column) and Yeh's method (4th column) supporting the wrongly predicted label and the most influential training point resisting the true label by our method(3rd column) and Yeh's method (5th column).

To start with, we train a bidirectional GAN to obtain the latent representation of data. The latent distribution $p_{\mathbf{z}}$ is set to be $[U_{[-1,1]}]^{100}$, a 100-D uniform distribution. Both encoder E and discriminator D consist of 3 convolutional layers with ReLU activations and 3 deconvolutional layers with ReLU activations are used for generator G. In Figure C.5, we present synthetic examples $G(z)$, real images x as well as the corresponding reconstructions $G(E(x))$. The qualitative results show that the generator and encoder are reliable in approximates data distribution and encoding data into latent representation.

C.2.1 Image exploring

In this experiment, we use a multi-layer perceptrons for classification, where one hidden layer with 512 units and ReLU activation function are adopted. Our method uses the latent representation

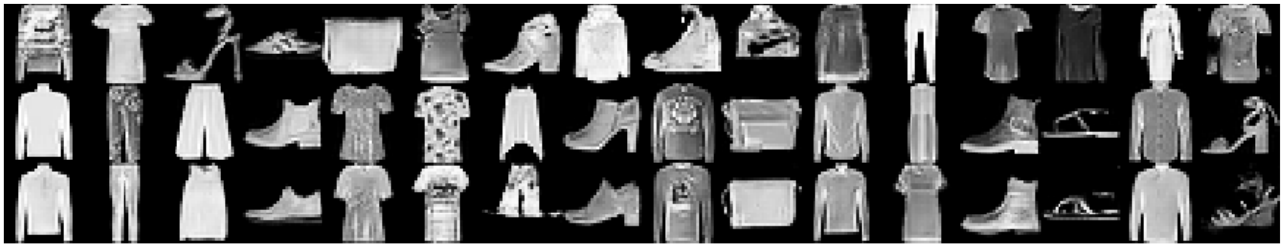


Figure C.5: BiGAN training results for Fashion-MNIST dataset,including generated samples $G(z)$(upper row), real data x(middle row), and corresponding reconstructions $G(E(x))$ (bottom row).

learned from BiGAN as model input. The L_2 weight decay is set to 0.003 for Yeh's and our methods for fair comparison. Both methods achieve test accuracy around 88%. We randomly select a few test points that are correctly classified by both methods and visualize three top training points with either positive or negative representer values.

In Figure C.6, a test sample of sandal, our method recovers top three positive images with label of sandal, which successfully capture the core patterns of the test image including curved top line, toe tip, as well as outsole. As for returned top three negative images, the first one has low heels, however, the outline of sandal still looks similar to the test image. For Yeh's method, it recovers the same rank 1st positive image as our method. However, the 2nd and 3rd positive images and all negative images are distant from the test image.

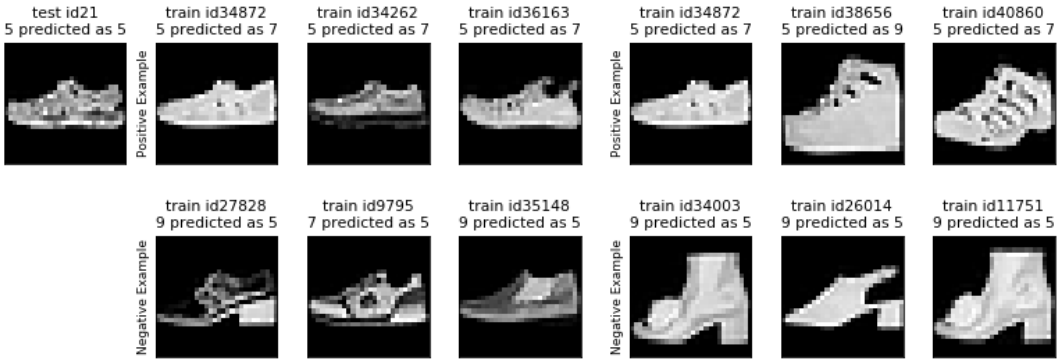


Figure C.6: Comparison of top three excitatory and inhibitory influential training images for a test point(ID 21) (left-most column) using our method (left columns) and Yeh's method (right columns).

Figure C.14 represents a set of recovered influential training points of a test image of coat. We notice that all recovered positive training images on the left columns have a fairly similar shape with the coat in test image. And surprisingly, the top 1st and 2nd images even try to capture the feature of white. neckline. We also find that negative images selected by our method look close to the test image and we can clearly observe a similar armhole. By contrast, training images recovered by Yeh's method are too obscure to tell the details of coats. More interestingly, sleeves and body are apart in positive images on right columns, which contradicts the test image. It is not surprising that our method outperforms Yeh's since our method measures influences of training samples on latent space, which focuses on similarity of features rather than Euclidean distance. More examples are included in the last section.

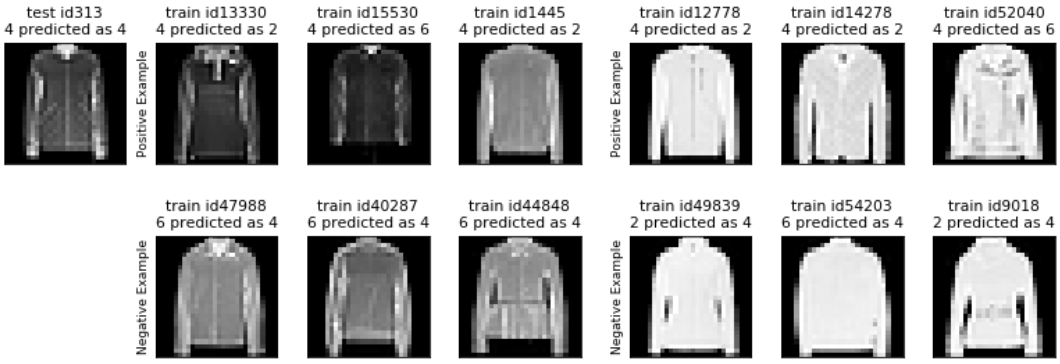


Figure C.7: Comparison of top three excitatory and inhibitory influential training images for a test point(ID 313) (left-most column) using our method (left columns) and Yeh's method (right columns).

C.2.2 Understanding misclassified images

Not only do we explore images both methods make correct prediction, we also want to investigate to what extent training points contribute to fooling the classifier. Indeed, two types of training points are of our interest, one of which try to fool model to stay away from true label and the other of which encourages model to make the wrong prediction. We randomly select a set of test points that both models fail in. In the first row of Figure C.8, the first column represents images

misclassified by both models. The 2nd and 3rd column represent training points selected by our method, which most endorses the wrong prediction and most fools the model away from ground truth respectively. Similarly, the 3rd and 4th columns represents two types of points recovered by Yeh's method. It is not hard to tell that our method outperforms Yeh's method in all examples since the influential samples selected by our method look much more close to test images. In particular, in the 4th row of Figure C.8, a shirt is predicted as T-shirt. Training images recovered by our method have short sleeves, which have similar shape as the shirt. However, training images recovered by Yeh are with long sleeves, which does not incorporate important features of the shirt at all. More examples can be found in the last section.

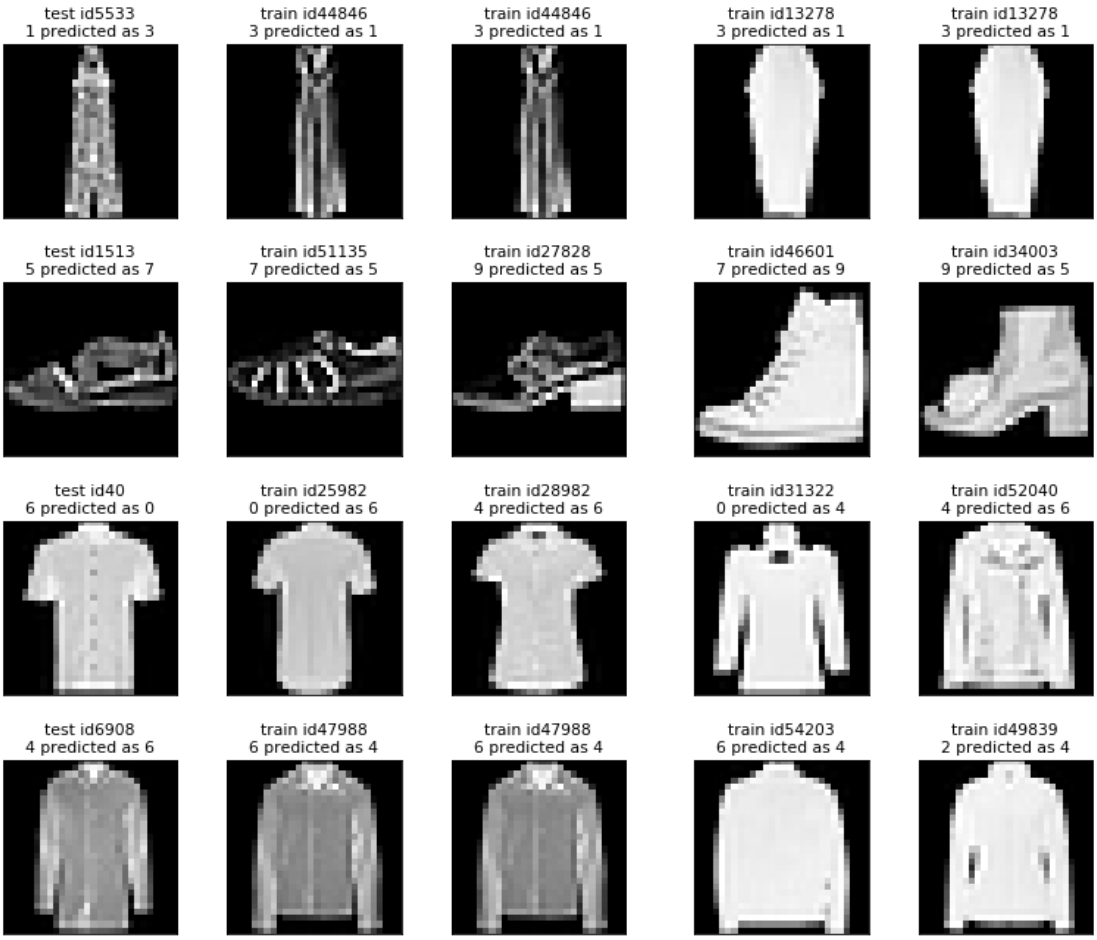


Figure C.8: A misclassified test image (left most) and the most influential training point by our method(2nd column) and Yeh's method (4th column) supporting the wrongly predicted label and the most influential training point resisting the true label by our method(3rd column) and Yeh's method (5th column).

C.2.3 More examples

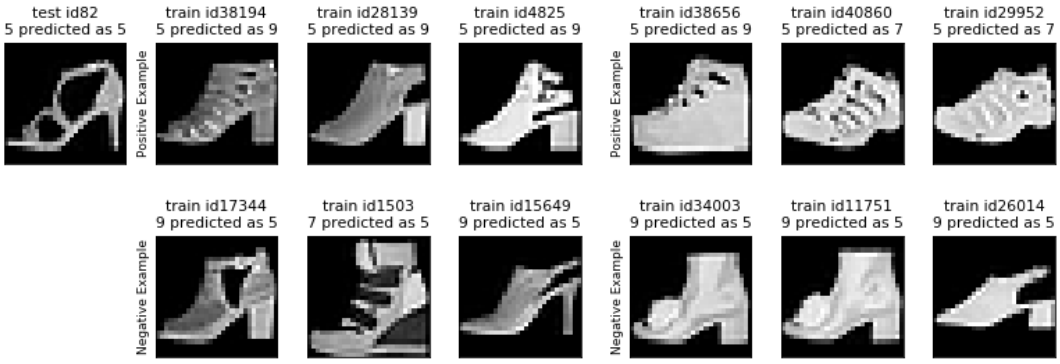


Figure C.9: Comparison of top three excitatory and inhibitory influential training images for a test point(ID 82) (left-most column) using our method (left columns) and Yeh's method (right columns).

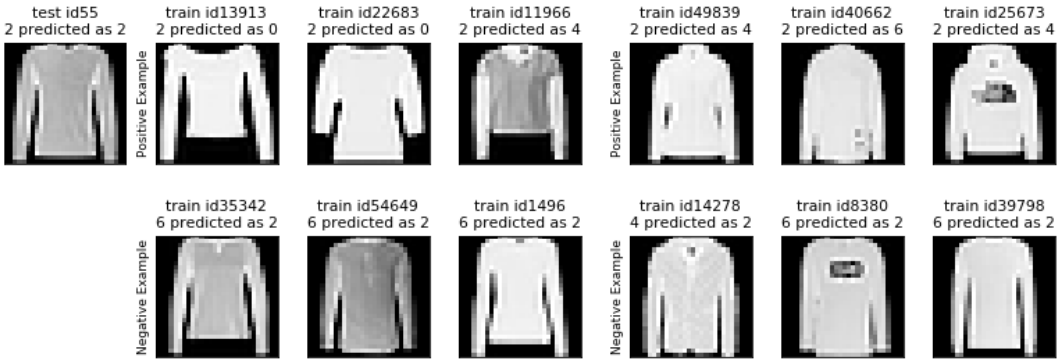


Figure C.10: Comparison of top three excitatory and inhibitory influential training images for a test point(ID 55) (left-most column) using our method (left columns) and Yeh's method (right columns).

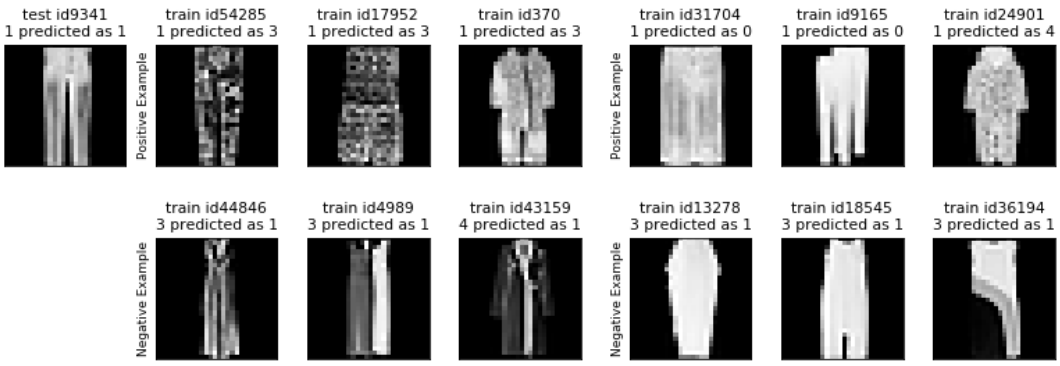


Figure C.11: Comparison of top three excitatory and inhibitory influential training images for a test point(ID 9341) (left-most column) using our method (left columns) and Yeh's method (right columns).

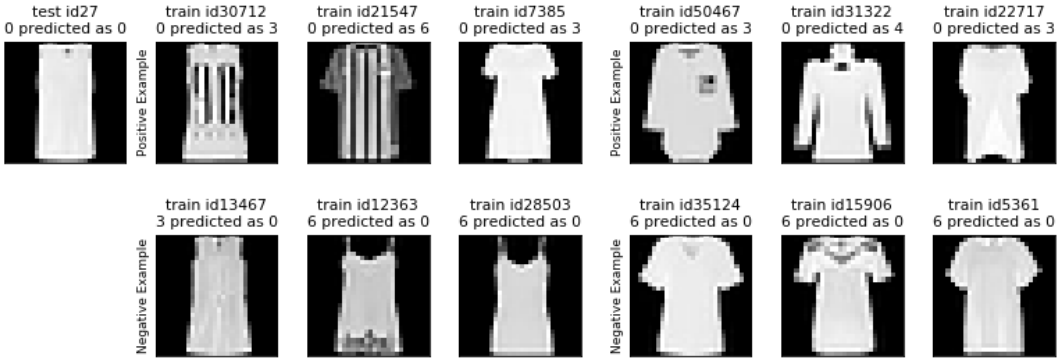


Figure C.12: Comparison of top three excitatory and inhibitory influential training images for a test point(ID 27) (left-most column) using our method (left columns) and Yeh's method (right columns).

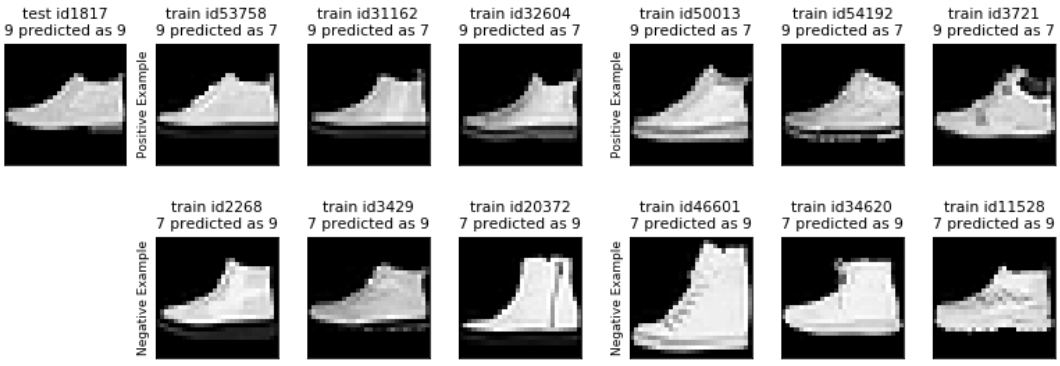


Figure C.13: Comparison of top three excitatory and inhibitory influential training images for a test point(ID 1817) (left-most column) using our method (left columns) and Yeh's method (right columns).

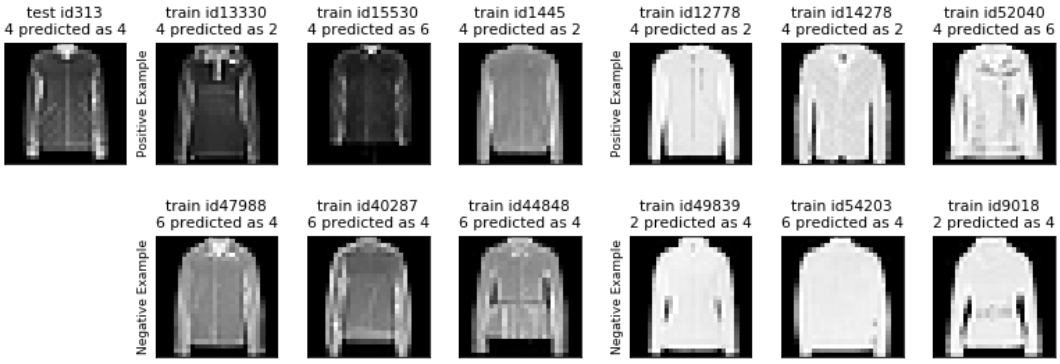


Figure C.14: Comparison of top three excitatory and inhibitory influential training images for a test point(ID 313) (left-most column) using our method (left columns) and Yeh's method (right columns).

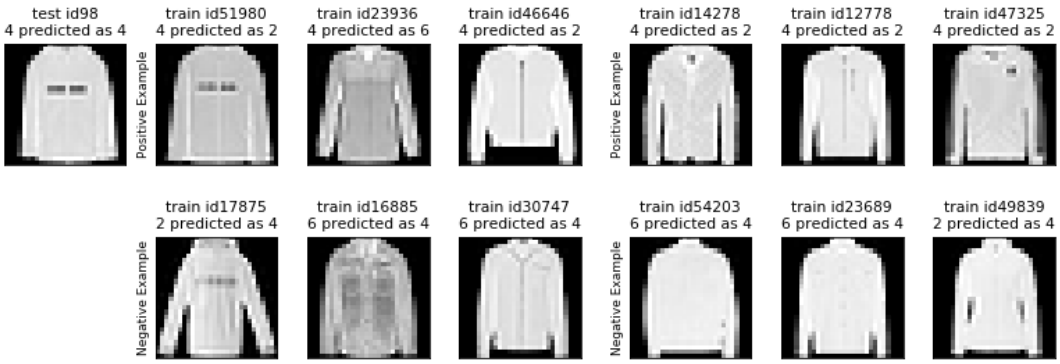


Figure C.15: Comparison of top three excitatory and inhibitory influential training images for a test point(ID 98) (left-most column) using our method (left columns) and Yeh's method (right columns).

Figure C.16: A misclassified test image (left most) and the most influential training point by our method(2nd column) and Yeh's method (4th column) towards the wrongly predicted label and the most influential training point resisting the true label by our method(3rd column) and Yeh's method (5th column).

Figure C.17: A misclassified test image (left most) and the most influential training point by our method(2nd column) and Yeh's method (4th column) towards the wrongly predicted label and the most influential training point resisting the true label by our method(3rd column) and Yeh's method (5th column).