ABSTRACT

LUO, SHIKAI. Optimal Individualized Treatment Strategy: Flexible Modeling with/without Imaging Covariates. (Under the direction of Rui Song and Subhashis Ghoshal.)

The personalized medicine has recently received increasing attention in both academia and industry. Different people may respond differently to the same treatment and the same person may respond differently to different treatments as well. Different from the standard treatment discovery framework which is used for finding a single treatment for a homogenous group of patients, personalized medicine involves finding therapies that are tailored to each individual in a heterogeneous group. A treatment regime (strategy) is a function that maps personalized charateristics to available treatment decisions. Using existing data with or without patients' imaging features, a key goal is to identify the optimal treatment strategy among all feasible ones, which can provide guidance to physicians and help patients to achieve the most favorable clinical outcome on average. In this thesis, we focus on estimation of the optimal treatment strategy under a new semiparametric additive single-index model framework and when patients are associated with image features.

In Chapter 2, we propose a new semiparametric additive single-index model for estimating individualized treatment strategy. The model assumes a flexible and nonparametric link function for the interaction between treatment and predictive covariates, while the optimal decision rule is determined by a simple linear combination of the covariates. We estimate the rule via monotone B-splines and establish the asymptotic properties of the estimators. Both simulations and an real data application demonstrate that the proposed method has a competitive performance.

In Chapter 3, we propose two novel approaches to estimate the optimal treatment strategy that uses both scalar and imaging covariates. The first approach assumes that the slope function of the imaging covariates belongs to the space of bounded total variation in order to account for the smooth nature of most imaging data. We develop an efficient penalized total variation

optimization to estimate the unknown slope function and other related parameters. We also establish the error bounds for the total variation slope estimator of imaging covariates and the coefficients of scalar covariates. The second approach is built upon convolutional neural network (CNN) which exploits the correlation between adjacent pixels in the two or three dimensional imaging space. We take this opportunity to employ deep learning to approximate the contrast function and assign future patients according to the sign of the estimated contrast function. Extensive simulations demonstrate that the two proposed methods have superior performance against other possible approaches.

Optimal Individualized Treatment Strategy: Flexible Modeling with/without Imaging Covariates

by
Shikai Luo

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Statistics

Raleigh, North Carolina

2016

APPROVED BY:

Yichao Wu	Hongtu Zhu
Rui Song Co-chair of Advisory Committee	Subhashis Ghoshal Co-chair of Advisory Committee

DEDICATION

To my beloved family.

BIOGRAPHY

The author was born in 1989 at a beautiful small town in Chongqing, China. In 2007, he was admitted by Shiing-Shen Chern mathematics class at Nankai University. He met many life-long friends and established his advanced mathematical skills. He found his love in solving difficult real-world problems with his math skills. After receiving the Bachelor's degree of mathematics in 2011, he attended North Carolina State University to pursue a PhD in statistics. With the valuable instruction and guidance from his advisors Dr. Rui Song, Dr. Subhashis Ghoshal and committee member Dr. Hongtu Zhu, he has obtained comprehensive research training, and a wide range of exposure to both statistical methodology and application. He will complete his PhD in May 2015 and join Quantlab Financial as a quantitative research scientist.

ACKNOWLEDGEMENTS

First of all, I would like to express my sincere gratitude to my advisors Dr. Rui Song and Dr. Subhashis Ghoshal for their continued support and great mentoring. Their patience, inspiring ideas and extraordinary kindness never fail to surprise me. They have offered invaluable support for my study and career development. I have learned a lot from Dr. Song and Dr. Ghoshal, not only on how to conduct research, but also the passion towards truth. It is a great experience to work with them!

I would also like to extend my appreciation to my committee members, Dr. Yichao Wu and Dr. Hongtu Zhu, for their constant encouragement and insightful comments. In particular, I want to express my sincere gratitude to Dr. Hongtu Zhu who treats me like his own son. He has taught me lots of technical skills and showed me how to find true happiness. I also thank Dr. Min Kang for teaching me stochastic analysis and kindly serving on my committee as the graduate school representative.

Next, I would like to thank all the wonderful professors at NC State. I am especially thankful to Dr. Hua Zhou and Dr. Eric Laber for teaching advanced computing course and stimulating me to learn lots of computing skills in my spare time. I also thank Dr. Sujit Ghosh, Dr. Howard Bondell, and Dr. Donald E. K. Martin who served as the DGPs providing invaluable suggestions and are always there for help. I also want to thank Terry Byron, Chris Waddell for helping me with computing problems and Alison McCoy, Arian Blue and Lanakila Alexander for their great service to the department. Especially, I want to thank Dr. Charles Smith for his knock, knock with snacks:-)

I also owe my gratitude to all my mentors during the two internships. In particular, Dr. Xiao Ni and Dr. Bill Prucka at Eli Lilly and Company, taught me not only lots of statistical ideas and professional skills for clinical trials, but also how to communicate with other collegues at industry. I want to thank Dr. Andrei Prudius and Dr. Frederick Tong-Uk Lee at Bloomberg for showing me how to apply quantitative skills to solve challenging financial problems.

Thank all my friends and fellow students at NC State. I would like to thank Dehan, Xiaoshan, Peng, Meng, Wenhao and many others, thank you! You are my family too!

Last, I want to express my deepest love to my family. I like to thank my father, who loved me so much. Dad, I love you and I wish you know that. To my great mother, your unconditional support and love is always the source of my strength. To my uncle, aunt, grandparents, and maternal grandparents, you love me so much and I love you, too!

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

There is an increasing interest in personalized medicine, which involves making treatment decisions for an individual patient using all information available on the patient, including genetic, demographic and other clinical variables, to achieve the best outcome for the patient based on this information. The reason behind personalized medicine is that different people may respond differently to the same treatment and the same person may respond differently to different treatments. Compared to the standard treatment discovery framework which is used for finding treatments for a homogeneous group of patients, personalized medicine may provide more accurate and more effective treatments in a heterogeneous group. A treatment regime formalizes personalized medicine as a decision rule for choosing a treatment for each patient based on that individual's observed characteristics. The optimal treatment regime is the decision rule that would maximize the clinical outcome on average conditional on the patient's information (Assume that larger outome is better).

Formally, the data contains n subjects sampled from a population of interest, with independent and identically distributed observations $\{(Y_i, A_i, X_i)\}_{i=1}^n$, where Y_i is the observed outcome, A_i is the received treatment, and X_i are covariates. Assume larger values of Y are preferred, X contains 1 as its first element and $A \in \{-1, 1\}$ for simplicity. The decision rule or treatment regime, d, is a function that maps the patient charateristics, x, to the treatment,

$a \in \{-1, 1\}$. A patient with covariates $X = x$ would receive treatment -1 if $d(x) = -1$ and treatment 1 if $d(x) = 1$. We aim to find the optimal treatment regime, d^{opt}, which yields the maximum mean clinical outcome.

Three fundamental assumptions are required to make this feasible. The first one is the stable unit treatment value assumption (SUTVA) (Rubin, 1978), also known as consistency assumption, which states that an individual's observed outcome under treatment A is the same as the potential outcome for that assigned treatment regardless of the treatment assignment mechanism and there are no interference among treatment effects, which may be violated in infectious disease cases. That is,

$$Y = Y^*(1)\frac{1+A}{2} + Y^*(-1)\frac{1-A}{2}, \tag{1.1}$$

where $Y^*(a)$ is the potential outcome of a patient if he or she is assigned to treatment $A = a$. The second one is the strong ignorability assumption, also known as the no unmeasured confounders assumption (Rosenbaum and Rubin, 1983), which assumes the conditional independence between the potential outcome and the treatment assignment. That is,

$$\{Y^*(1), Y^*(-1) \perp\!\!\!\perp A|X\}. \tag{1.2}$$

The third one is the positivity assumption, stated as $0 < P(A = 1|X) < 1$ almost surely. Both the consistency assumption and the no unmeasured confounder assumption are not testable based on observed data.

Based on above three assumptions, people have developed lots of approaches to estimate the optimal treatment regime based on data from clinical trials or observational studies, which can involve a single decision point or a series of sequential decision points. We only consider studies with one single decision point. A backward recursive fitting procedure related to dynamic programming algorithm (Bather, 2000) can be implemented to deal with studies with multiple decision points. In order to find the optimal treatment regime, an important issue is how to

evaluate a given regime d, or equivalently, how to compute the mean potential outcome $E\left[Y^*(d)\right]$ if the population of interest follows this regime. It is straightforward to show that

$$E\left[Y^*(d)\right] = E_X\left[E\left[Y^*(1)|X\right]\frac{1+d(X)}{2} + E\left[Y^*(-1)|X\right]\frac{1-d(X)}{2}\right], \tag{1.3}$$

where we can introduce the popular Q-function as

$$Q(x,a) = E\left[Y^*(a)|X=x\right] = E\left[Y|X=x, A=a\right], \tag{1.4}$$

and the optimal treatment regime is given by

$$d^{opt}(x) = \text{sign}\{Q(x,1) - Q(x,-1)\}, \tag{1.5}$$

where we can further define the contrast function as

$$C(x) = \left[Q(x,1) - Q(x,-1)\right]/2. \tag{1.6}$$

The above simple derivation has led to two popular regression-based approaches for estimating the optimal treatment regime, Q-learning (Q denoting "quality") (Watkins and Dayan, 1992; Murphy, 2005; Zhao et al., 2009, 2011; Moodie et al., 2014; Song et al., 2015b) and A-learning (A denoting "advantage") (Murphy, 2003). Q-learning uses a posited regression model for the outcome of interest given patient's covariates information at each decision point, the so-called Q-function given at (1.4). It turns out that the optimal treatment regime depends only on the contrast function (1.6). This contrast function represents the contrasts among treatments and is exactly the only part of the regression outcome that needs to be explicitly modeled in A-learning. An advantage of Q-learning is that it incurs limited computational burden, however, Q-learning can be sensitive to misspecification of the require models, while A-learning enjoys the so-called double robustness property in that the corresponding estimating equations are asymptotically unbiased when either the propensity score or the Q-function (baseline function

if we assume the contrast is already correctly specified) is correctly specified.

The Q-learning has shown to converge to the optimal treatment regime with probability one with discrete covariates and treatment variables. Murphy (2005) derived an upper bound on the generalization error for the Q-learning due to model approximation. Zhao et al. (2009) used support vector regression models to mitigate the risk of mismodeling issue. Qian and Murphy (2011) suggested to deal with high dimensional covariates via ℓ_1 penalty and established an error bound. Song et al. (2015b) proposed the penalized Q-learning, which enables valid inference and reduces computational burden. However, the penalized Q-learning can only handle discrete covariates. Goldberg et al. (2013) extended the penalized Q-learning to continuous covariates, called adaptive Q-learning. In addition, Chakraborty et al. (2013) proposed an adaptive m-out-of-n bootstrap to construct confidence intervals for parameters indexing the optimal regime. Laber et al. (2014a) proposed the interactive Q-learning which avoids the non-smooth and non-monotone Q-functions and establishes consistency under more general models than the standard Q-learning procedure. Developments of statistical learning suggest a large collection of flexible and powerful regression methods that can be used for modeling Q-functions, see Taylor et al. (2015); Craven and Shavlik (1996) for some discussion.

Murphy (2003) introduced a regret-function based A-learning. The regret function is the amount of improvement that could be made if the optimal treatment were employed. Other regret regression modeling versions of A-learning have been proposed by Henderson et al. (2010); Barrett et al. (2014). Robins (2004) proposed another version of A-learning knowns as the g-estimation, using structural nested mean models to target the contrast function explicitly. Vansteelandt et al. (2014) provided valuable discussion on structural nested mean model and the g-estimation. Moodie and Richardson (2010) proposed an approach called Zeroing Instead Plugging In, which gives nearly the same estimator to that of g-estimation and is referred as the hard-threshold estimator by Chakraborty et al. (2009). Moodie et al. (2007) explored the connection between the g-estimation and the regret modeling method, and showed that they are closely related but not equivalent. Barrett et al. (2014) proved that the doubly robust version

of the regret regression is equivalent to the g-estimation. In addition, Lu et al. (2011) developed a penalized regression framework to achieve variable selection within the A-learning procedure. Schulte et al. (2014) provided a detailed explanation of Q- and A-learning.

The optimal treatment regime can also be estimated via direct search based on propensity score. The propensity score is defined as

$$\pi(x) = P(A = 1|X = x), \tag{1.7}$$

which can be estimated by logistic regression or any other standard classifier. Zhang et al. (2012b) proposed an inverse probability weighted estimator to estimate the corresponding potential outcome $E\left[Y^*(d)\right]$ for a given regime $d(x;\eta)$ indexed by a finite dimensional parameter η. Formally, the inverse probability weighted estimator is $n^{-1}\sum_{i=1}^{n} w(X_i;\eta)Y_i$ where the weight $w(X_i;\eta)$ is given by

$$w(X;\eta) = \mathbf{1}\{A = d(X;\eta)\}/P(A = d(X;\eta)|X). \tag{1.8}$$

They also proposed a doubly robust and more efficient augmented version of the inverse probability weighted estimator, given as

$$n^{-1}\sum_{i=1}^{n} w(X_i;\eta)Y_i + (w(X_i;\eta) - 1)\left[Q(X_i, 1)\frac{1 + d(X_i;\eta)}{2} + Q(X_i, -1)\frac{1 - d(X_i;\eta)}{2}\right]. \tag{1.9}$$

The optimal treatment regime is directly searched within a group of regimes in the form of $\text{sign}\{\eta^T x\}$. Zhang et al. (2013) extended above method to the estimation of optimal dynamic treatment regime. Zhang et al. (2012a) proposed to estimate the optimal treatment regime within the framework of classification where the optimal classfier is the optimal regime. Song et al. (2015a) also took the treatment regimes as classifier and proposed the penalized outcome weighted learning to deal with high dimensional covariates. They established variable selection consistency property and the asymptotic distribution of related estimators.

The contribution of this thesis is as follows. First, we proposed a semiparametric single-index model to estimate the optimal treatment regime which is both flexible and interpretable. Our model provides a more flexible interaction between the covariates and the treatment compared to traditional parametric models, in which we allow a fully nonparametric baseline function and a close-to nonparametric contrast function characterizing the interaction of the treatment A and the covariates X. Within this framework, the optimal treatment regime is simply the sign function of a linear combination of covariates X.

Second, while lots of researchers have focused on methods for making treatment decisions based on scalar covariates at one time point or a sequence of time points, decisions based on both scalars and images have much remained unexplored (McKeague and Qian, 2014; Ciarleglio et al., 2015). McKeague and Qian (2014) proposed to estimate and evaluate the treatment regimes based on one baseline functional covariate. Ciarleglio et al. (2015) used both scalars and multiple functional covariates to select the optimal treatment. However, neither has directly dealt with two or three dimensional imaging features. We develop two approaches to estimate the optimal treatment regime directly operating on patients' imaging features. The first approach assumes that the contrast function is a simple linear function of both scalars and imaging features. Furthermore, the slope function of the image covariates belongs to the space of bounded variation in order to account for the piecewise smooth nature of most imaging data. We develop an efficient penalized total variation optimization algorithm to estimate the unknown slope function and other related parameters and establish the error bounds for the total variation slope estimator. Furthermore, when the contrast function is not believed to be linear, we employ a convolutional neural network (CNN) to directly learn the nonlinear contrast between the two treatment groups and assign future patients according to the sign of the estimated contrast function. The performance of our proposed two approaches are investigated through extensive simulations.

Chapter 2

Semiparametric Single-Index Model for Estimating Optimal Treatment Strategy

In modern clinical researches, the goal to achieve better outcomes as well as lower cost and burden for individual patients has generated tremendous interest in personalized medicine. Individualized treatment rules (ITRs) operationalize personalized medicine as a decision function from patient's individual biomarkers to a recommended treatment and the optimal ITRs should be the one which maximizes clinical benefit if implemented. Specifically, if we use A to denote treatment assignment taking values of -1 and 1, X to denote all biomarker and prognostic information associated with each patient and let Y be the clinical outcome of interest (assuming large values are desirable), then an individualized treatment rule (ITR), denoted by $d(x)$, takes a given value x of X and provides a treatment choice from $\{-1, 1\}$. Furthermore, let P^d denote the distribution of (X, A, Y) and expectation with respect to this distribution by E^d, where the individualized treatment rule $d(x)$ is used to assign treatments. Define the value function as $V(d) = E^d(Y)$. Then an optimal ITR, d^{opt}, is a rule that has the maximal value, i.e., d^{opt} is the maximizer of $V(d)$ over decision rules d.

There has been growing interest in developing valid inference methods for estimating the optimal ITRs, d^{opt}, using clinical trial data. With trial data, it holds $V(d) = E[YI(A = d(X))/\pi(A|X)]$ (Qian and Murphy, 2011), where $\pi(a|X)$ is the known randomization probability of $A = a$ given X, so it is easy to see $d^{opt}(x) = \text{sign}\{E[Y|A = 1, X = x] - E[Y|A = -1, X = x]\}$, where sign$(\cdot)$ function is defined as sign$(x) = 1$ when $x > 0$, sign$(x) = -1$ when $x < 0$. Therefore, most of the existing methods tend to model $E[Y|A = a, X = x]$ including the interactions between the treatment and the covariates either parametrically or nonparametrically. Such literature include likelihood-based approach (Thall et al., 2000, 2002, 2007), parametric Q-learning in Chakraborty et al. (2010), and machine learning based methods (Zhao et al., 2011). Alternatively, one can parametrically model $E[Y|A = a, X = x] - E[Y|A = d^{opt}(X), X = x]$ which is called A-learning as discussed in Murphy (2003) and Robins (2004). Recently, directly maximizing $V(d)$ has been proposed using support vector machine in Zhao et al. (2012) or via robust parametric models in Zhang et al. (2012b). However, all parametric methods potentially suffer from model misspecification especially when X is not low-dimensional and the optimal ITRs depends on high-order interactions among X's. On the other hand, although the nonparametric methods such as machine-learning methods are flexible, the resulting rules are complicated so may not be interpretable in practice. The latter often comes with no rigorous inference procedures as in the parametric methods.

In this chapter, we propose a semiparametric single-index model to estimate the optimal ITRs. Our model retains a flexible and nonparametric formulation of the treatment-covariate interactions but also yields a simple decision rule which only depends on a linear combination of X. Specifically, our proposal assumes the following model between Y and (X, A):

$$E[Y|X, A] = \mu(X) + \psi(\beta^T X)A, \tag{1}$$

where X is a p-dimensional covariate vector and may contain 1 as the intercept, $\beta^T X$ is a single index and both μ and ψ are unknown functions. Moreover, ψ is a monotone increasing function

with $\psi(0) = 0$. The proposed model has the following advantages in developing individualized treatment strategy. First, it provides a more flexible interaction between the covariates and the treatment as compared to the traditional parametric models, in which we allow a fully nonparametric baseline function of the covariates X, $\mu(X)$, and a close-to nonparametric interaction between the treatment A and the covariates X. Second, more importantly, under this model, since ψ is increasing and $\psi(0) = 0$, we can easily derive the best treatment strategy as a very simple rule:

$$d^{opt} : X \longrightarrow \text{sign}(\beta^T X).$$

Therefore, the resulting rule is practically interpretable and useful as only the sign of a risk score $\beta^T X$ needs to be evaluated for each patient. As a separate note, single index models have been studies extensively in literature with a number of inference methods developed, including the average derivative method (Horowitz and Härdle, 1996), the sliced inverse regression Li and Duan (1989); Duan and Li (1991); Li (1991), the iterative average derivative method (Hristache et al., 2001) and other related methods (Xia, 2006). Estimating both the single index and the link function at the same time has also been studied in (Klein and Spady, 1993; Ichimura, 1993; Hardle et al., 1993). However, none of these works have considered the single index model for estimating the optimal ITRs, especially that our model (1) assumes the main effect of X, $\mu(X)$, to be fully nonparametric.

The rest of the chapter is organized as follows. In Section 2, we provide a full inference procedure for the proposed semiparametric single index model. Extensive simulation studies are presented in Section 3 and a real data analysis is presented in Section 4, followed by a discussion section.

2.1 Inference Procedure

Note that model (1) remains the same if we replace $\psi(x)$ by $\psi(rx)$ for any $r > 0$. Therefore, for identifiability, we further require $\|\beta\| = 1$ where $\|\cdot\|$ is the Euclidean ℓ_2-norm in R^p. Assume

that data are obtained from a randomized trial with i.i.d observations (Y_i, X_i, A_i), $i = 1, ..., n$. The randomization probability $P(A = a|X) = \pi(a|X)$ is known by the trial design.

To avoid estimating the nonparametric function $\mu(X)$ when making inference for β, we first observe that,

$$E\left[\frac{AY}{2\pi(A|X)}|X\right] = E\left[\frac{A}{2\pi(A|X)}E\left[Y|A,X\right]|X\right]$$
$$= E\left[Y|A=1,X\right]/2 - E\left[Y|A=-1,X\right]/2 = \psi(\beta^T X).$$

Therefore, a natural estimate of β is obtained by minimizing the least square, given as

$$\sum_{i=1}^{n}\left\{\frac{A_i Y_i}{2\pi(A_i|X_i)} - \psi(\beta^T X_i)\right\}^2,$$

subject to $\|\beta\| = 1$. Since ψ is an increasing function, we approximate $\psi(x)$ using monotone B-spline basis, which are indicated as $N_1(x), ..., N_{K_n+M}(x)$, where K_n is the number of interior knots with equal partition in an interval containing $\beta^T X$ and M is B-spline order, i.e., for cubic B-spline, $M = 4$. Additionally, we impose an upper bound M_n for the summation of absolute values of all the B-spline coefficients to prevent divergence of these coefficients in the minimization. M_n is a constant depending on n and the choice of M_n is discussed in Section 3. Thus, the minimization becomes

$$\begin{aligned}
\min_{\xi,\beta} \quad & \sum_{i=1}^{n}\left\{\frac{A_i Y_i}{2\pi(A_i|X_i)} - \sum_{j=1}^{K_n+M}\xi_j N_j(\beta^T X_i)\right\}^2, \\
\text{subject to} \quad & \|\beta\| = 1,\ \xi_1 \le \cdots \le \xi_{K_n+M},\ \sum_{j=1}^{K_n+M}|\xi_j| \le M_n,\ \sum_{j=1}^{K_n+M}\xi_j N_j(0) = 0. \quad (2.1)
\end{aligned}$$

Set $d = K_n + M$. The objective function in (2.1) is quadratic in ξ and quite nonlinear in β. The constraint $\|\beta\| = 1$ is nonlinear in the elements of β. The inequality constraint in (2.1) is linear in ξ since it can be expressed as $B\xi \le 0$, where $\xi = (\xi_1, \cdots, \xi_d)^T$ and B is a $(d-1) \times d$ matrix with $B(i,i) = 1, B(i,i+1) = -1$ and the rest of its entries being zero. To facilitate the

implementation, we now propose an iterative estimation algorithm to solve (2.1). In particular, we iteratively solve β with ξ fixed at their current values, and then solve ξ with β fixed at their current values, and repeat them until the convergence criterion is met. The computation procedure can be summarized as the following.

Step 1: Get an initial estimator $\widehat{\beta}^{(0)}$. For example, we can set $N_j(\beta^T X) = \beta^T X$ as a linear function in (2.1) and compute the ordinary least squares (OLS) estimator for β. Normalize $\beta^{(0)}$ such that $\|\beta^{(0)}\| = 1$. Set $k = 0$.

Step 2: Given the initial estimates of the index values $\{Z_i = \widehat{\beta}^{(k)T} X_i, i = 1, \cdots, n\}$, minimize over ξ by solving the followng quadratic programming (QP) problem:

$$\begin{aligned} \min_{\xi} \quad & Q(\xi) = \sum_{i=1}^{n} \left\{ \frac{A_i Y_i}{2\pi(A_i|X_i)} - \sum_{j=1}^{d} \xi_j N_j(Z_i) \right\}^2, \\ \text{subject to} \quad & B\xi \leq 0 \text{ and } \sum_{j=1}^{K_n+M} |\xi_j| \leq M_n, \quad \sum_{j=1}^{K_n+M} \xi_j N_j(0) = 0. \end{aligned} \tag{2.2}$$

Denote the solution as $\widehat{\xi}^{(k)}$.

Step 3: Fix ξ at the current values, minimize

$$\sum_{i=1}^{n} \left\{ \frac{A_i Y_i}{2\pi(A_i|X_i)} - \sum_{j=1}^{d} \widehat{\xi}_j^{(k)} N_j(\beta^T X_i) \right\}^2, \quad \text{s.t.} \ \ \|\beta\| = 1.$$

Denote the solution as $\widehat{\beta}^{(k+1)}$. This problem can be solved using the nonlinear least squares (NLS) algorithm.

Step 4: Set $k = k + 1$. Go to Step 2 and iterate until convergence.

In our numerical examples, we use the MATLAB's optimization toolbox: the function quadprog() for QP in Step 2 and lsqnonlin() for NLS in Step 3.

Given K_n, we choose to place the interior knots at equally-spaced sample quantile of the predictor variable, which is $\beta^T X$ in this context. For example, if there are 4 interior knots, then they would be respectively at the 20th, 40th, 60th, 80th percentile. The boundary knots are naturally chosen as the minimum and maximum values of the predictor variable. During the iteration, the estimated single index β could change at each step, therefore the knots also change in the iteration. The number of knots K_n can be tuned with cross-validation. In general, 5 to 10 knots will be sufficient to have very good results.

2.2 Asymptotic Results

We establish the asymptotic properties of the estimators $(\widehat{\beta}_n, \widehat{\psi}_n)$, including their consistency under certain metric, the convergence rates, and the asymptotic distribution of $\sqrt{n}(\widehat{\beta}_n - \beta_0)$. We need the following conditions.

(C.1) β_0 is assumed to be in the unit ball $\mathcal{B}$ of R^p and X has a compact support. In addition, $E(XX^T|\beta_0^T X)$ is positive definite. and $E[X|\beta_0^T X = x]$ is kth continuously differentiable with bounded derivatives for some $k > 3$.

(C.2) ψ_0 has bounded kth derivative in an open interval containing the support of $\beta_0^T X$ for some $k > 3$; moreover, $\psi_0'(0) > 0$.

(C.3) $E[\psi_0(\beta_0^T X)|\beta^T X]$ is continuously differentiable in β and moreover,

$$E\left[\nabla E[\psi_0(\beta_0^T X)|\beta^T X]\Big|_{\beta=\beta_0}^{\otimes 2}\right] > 0.$$

Under these conditions, we first obtain the consistency and convergence rate of $(\widehat{\beta}_n, \widehat{\psi}_n)$.

Theorem 1. Under (C.1)-(C.3), we further assume $K_n = C_1 n^\gamma$ and $M_n = C_2 n^\tau$ for some positive constants C_1, C_2 with $\gamma > 0$, $\tau \geq 0$, and $11\gamma + 9\tau \leq 1$, $2\tau \leq (2k-5)\gamma$. Let $0 < \nu < 1/2$, then

$$\|\widehat{\beta}_n - \beta_0\|^2 + \|\widehat{\psi}_n - \psi_0\|^2_{L_2[a,b]} = o_p(n^{-1+\nu}) + O_p(n^{-2k\gamma}).$$

Furthermore,

$$\|\widehat{\psi}_n - \psi_0\|_{W^{1,\infty}[a,b]} = o_p(1),$$

where $W^{s,\infty}$ is the Sobolev space consisting of functions with bounded lth derivatives for any $l \leq s$. Furthermore, the Sobolev norm is defined as $\|\psi\|_{W^{1,\infty}[a,b]} = \max_{\alpha \leq 1} \|\psi^{(\alpha)}\|_{L^\infty[a,b]}$.

The asymptotic distribution of $\widehat{\beta}_n$ is stated in the following theorem.

Theorem 2. In addition to (C.1)-(C.3), we assume $K_n = C_1 n^\gamma$ and $M_n = C_2 n^\tau$ for some positive constants $C_1,\ C_2$ with $\gamma > 1/(4k-4),\ \tau \geq 0$ and $11\gamma + 9\tau \leq 1,\ 2\tau \leq (2k-5)\gamma$. Then $\sqrt{n}(\widehat{\beta}_n - \beta_0)$ converges in distribution to a mean-zero normal distribution with covariance $\Sigma_1^{-1}\Sigma_2\Sigma_1^{-1}$, where

$$\Sigma_1 = E\left[\psi_0'(\beta_0^T X)^2 X X^T\right]$$

and

$$\Sigma_2 = E\left\{\mathrm{Var}\left[\frac{AY}{2\pi(A|X)} \mid X\right]\psi_0'(\beta_0^T X)^2 X X^T\right\}.$$

Based on Theorem 2, a consistent estimator for the asymptotic covariance is given by $\widehat{\Sigma}_1^{-1}\widehat{\Sigma}_2\widehat{\Sigma}_1^{-1}$ in which $\widehat{\Sigma}_1$ and $\widehat{\Sigma}_2$ are given as follows. Then an estimator for Σ_1 is given as

$$\widehat{\Sigma}_1 = n^{-1}\sum_{i=1}^n \widehat{\psi}_n'(\widehat{\beta}_n^T X_i)^2 X_i X_i^T.$$

Since

$$\Sigma_2 = E\left\{\left[\frac{AY}{2\pi(A|X)} - \psi_0(\beta_0^T X)\right]^2 \psi_0'(\beta_0^T X)^2 X X^T\right\},$$

an estimator for Σ_2 is given by

$$\widehat{\Sigma}_2 = n^{-1}\sum_{i=1}^n \left[\frac{A_i Y_i}{2\pi(A_i|X_i)} - \widehat{\psi}_n(\widehat{\beta}_n^T X_i)\right]^2 \widehat{\psi}_n'(\widehat{\beta}_n^T X_i)^2 X_i X_i^T.$$

Under Theorem 1, it is clear that both $\widehat{\Sigma}_1$ and $\widehat{\Sigma}_2$ are consistent estimators for Σ_1 and Σ_2 respectively when the sample size converges to infinity. Finally, we estimate the optimal decision rule as $\mathrm{sign}(\widehat{\beta}_n^T X)$. Under such a rule, for any subject, the reward gain of using the optimal rule

vs the non-optimal rule is estimated to be $2\mathbb{P}_n\left[|\widehat{\psi}_n(\widehat{\beta}_n^T X)|\right]$.

2.3 Simulation Studies

In this section, we conduct simulations to investigate the empirical performance of our proposed method and compare with the inverse probability weighted estimator(IPWE) and augmented inverse probability weighted estimator(AIPWE) in Zhang et al. (2012b).

We consider the model $Y = \mu(X) + \psi(\beta^T X)A + \epsilon$ where X is generated uniformly from $[-1, 1]^p$, A is generated as -1 and 1 with equal probability 0.5 and the noise ϵ follows a normal distribution with mean 0 and standard deviation $\sigma = 0.5$. The three examples are:

- Example I: $p = 2, \mu(X) = X_1X_2 + X_2^2, \psi(\beta^T X) = 2(\beta^T X)^3, \beta = \frac{1}{\sqrt{2}}(1, -1)^T$.
- Example II: $p = 3, \mu(X) = X_1^2 + 2X_1X_2, \psi(\beta^T X) = 0.6\exp(\beta^T X) - 0.6, \beta = \frac{1}{\sqrt{2.9}}(1, -0.95, 1)^T$.
- Example III: $p = 5, \mu(X) = X_1X_2 + X_3^2, \psi(\beta^T X) = (\beta^T X)^3, \beta = \frac{1}{\sqrt{4.8}}(1, -0.95, 1, -0.95, 1)^T$.

To evaluate the estimation performance of the single index coefficient, we report its bias and the mean squared error MSE(β) = average over replications of $\|\widehat{\beta} - \beta\|^2/p$. To evaluate the estimation performance of the link function, we report its mean squared error MSE(ψ) = average over replications of $\frac{1}{n}\sum_{i=1}^n \|\widehat{\psi}(\widehat{\beta}^T X_i) - \psi(\beta^T X_i)\|^2$. To evaluate the accuracy of a treatment assignment rule sign$(\widehat{\beta}^T X)$, we calculate the percentage of making correct decisions (PCD), i.e. $1 - \frac{1}{2n}\sum_{i=1}^n |\text{sign}(\widehat{\beta}^T X_i) - \text{sign}(\beta^T X_i)|$. We also study the behavior of the value function estimates. Based on the estimated rule, the value function can be estimated as $\frac{1}{n}\sum_{i=1}^n \frac{Y_i 1(A_i = g_i)}{P(A_i = g_i|X_i)}$, where g_i is the estimated rule. We compare the proposed method with Zhang et al. (2012b) in terms of parameter estimates, percentage of making correct decisions (PCD) and value function estimates.

From Tables 1–3, we observe that our method shows better results compared with the inverse probability weighted estimator (IPWE) and the augmented inverse probability weighted estimator (AIPWE) (Zhang et al., 2012b) in terms of smaller bias of estimated single index

coefficient, smaller mean square error of estimated link function. In most cases, the bias of estimated single index coefficient of our proposed approach is about ten times smaller than the other two approaches. As a result, our method also makes more correct decisions and gives estimated value function much closer to its theoretical value. We also note that as sample size increases, the mean squared error of the single index coefficient and estimated link function for three methods decreases, the PCD increases and the estimated value function gets closer to the true value function.

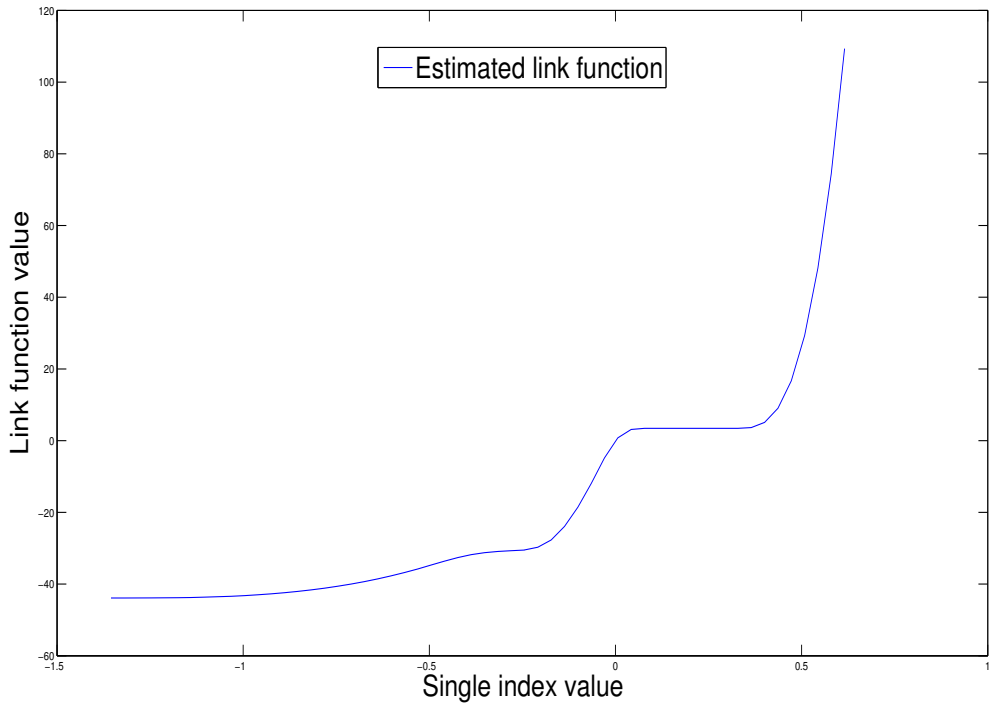
We also investigate our proposed inferential procedure for the single index coefficient β. It shows in Table 4 that, as sample size increases, the empirical standard error and the mean estimated standard error are getting closer to each other. For almost all cases, the empirical coverage rates are very close to the nominal level, as expected.

2.4 Real data analysis

To further illustrate the performance of our method, we consider its application to data from AIDS Clinical Trials Group Protocol 175 (ACTG175). The complete data contain 2139 HIV-infected subjects with study subjects randomized to four different treatment groups: zidovudine (ZDV) monotherapy, ZDV + didanosine (ddI), ZDV + zalcitabine and ddI monotherapy. The CD4 count (cells/mm^3) at 20 ± 5 weeks post-baseline is chosen as the continuous response Y, where large values are desired. Among all subjects, 524 subjects received the treatments ZDV + didanosine (ddI) and 522 subjects received the treatment ZDV + zalcitabine. For illustration purpose, we consider these two group of patients with the goal to find their individualized optimal treatment rules. We use $A = 1$ to denote treatment ZDV + zalcitabine and $A = -1$ to denote treatment ZDV + didanosine (ddI). Besides the treatment indicator, we also include two covariates: age and homosexual activity (in short as homo), which are selected as important covariates in Lu et al. (2011).

We apply the proposed method to estimate the optimal treatment and perform statistical inference for the corresponding parameters. The estimates for the single index coefficients

are 0.902, -0.036, and 0.430 respectively and the estimated variance of the single index coefficients are 0.2232, 0.0004 and 0.0984, respectively. The optimal treatment rule is sign(0.902-0.036×age+0.430×homo). That is, if 0.902-0.036×age+0.430×homo ≥ 0, the optimal treatment for this patient is ZDV + zalcitabine, otherwise, the optimal treatment is ZDV + didanosine(ddI). In other words, for a patient with homo = 0, the optimal treatment $A = -1$ if age > 25.2 and the optimal treatment $A = 1$ otherwise; while for a patient with homo = 1, the optimal treatment $A = -1$ if age > 37.2 and the optimal treatment $A = 1$ otherwise. We note that the age of study subjects ranges from 12 to 70. According to the estimated optimal rule, 565 out of 1046 patients (54.02%) in this subset should be assigned to treatment ZDV+didanosine (ddI).



2.5 Discussion

In this chapter, we proposed a novel semiparametric single-index model for individualized treatment selection. Our model plays an important role as a compromise between parametric models and nonparametric models (Zhang et al., 2012b). The decision rule based on our method is a simple linear combination of covariates. We provide statistical inference for this rule. The asymptotic properties for the proposed method are established. The proposed method demonstrates superior numerical behavior in terms of smaller bias and means square error. Based on the estimated rule, our method also provides more precise decisions than existing methods and gives more precise value function estimates.

In many clinical studies, the state space is often of very high dimension. To develop optimal individualized treatment rules in this case, it will be important to develop simultaneous variable selection and treatment rule estimation. Variable selection techniques such as penalized regression and variable screening can be nested into our semiparametric single index modeling framework as powerful tools to develop optimal individualized treatment rules.

In our current procedure, we assume the propensity score $\pi(A|X)$ is known. In observational studies, the propensity scores are often unknown. For such observational data, we can estimate $\pi(A|X)$ via logistic regression and plug-in the estimated propensity score funtion $\widehat{\pi}(A|X)$ into the optimization equation (2.1). It is beyond the scope of the current work and is an interesting topic for future study.

Table 2.1 Estimation and classification results for Example I. PCD denotes percentage of correct decisions, Val denotes value function estimates based on large sample. We report mean of estimated single index coefficient biases, mean squared errors of estimated single index coefficients, mean squared errors of estimated link functions, PCD and Val over 400 replications with their empirical standard errors one line below.

Method	Bias of (β_1, β_2)		MSE(β)	MSE (ψ)	PCD	Val(0.902)
	$n = 500$					
SIM	-0.001	0.000	0.001	0.008	0.986	0.900
	(0.001)	(0.001)	(0.000)	(0.000)	(0.001)	(0.004)
IPWE	0.016	0.086	0.055		0.863	0.951
	(0.007)	(0.009)	(0.004)		(0.003)	(0.004)
AIPWE	0.008	0.074	0.055		0.867	0.948
	(0.007)	(0.008)	(0.004)		(0.003)	(0.004)
	$n = 1000$					
SIM	0.000	0.000	0.000	0.004	0.990	0.902
	(0.001)	(0.001)	(0.000)	(0.000)	(0.000)	(0.003)
IPWE	0.001	0.054	0.045		0.883	0.939
	(0.007)	(0.008)	(0.003)		(0.003)	(0.003)
AIPWE	-0.007	0.041	0.045		0.885	0.937
	(0.007)	(0.007)	(0.003)		(0.002)	(0.003)
	$n = 1500$					
SIM	0.000	0.000	0.000	0.003	0.992	0.897
	(0.001)	(0.001)	(0.000)	(0.000)	(0.000)	(0.002)
IPWE	0.012	0.064	0.046		0.883	0.927
	(0.007)	(0.008)	(0.003)		(0.002)	(0.002)
AIPWE	0.004	0.053	0.046		0.885	0.925
	(0.007)	(0.007)	(0.003)		(0.003)	(0.002)

Table 2.2 Estimation and classification results for Example II. Other captions are the same as Table 1.

Method	Bias of $(\beta_1, \beta_2, \beta_3)$			MSE(β)	MSE (ψ)	PCD	Val(0.645)
	$n = 500$						
SIM	-0.011	0.008	-0.004	0.009	0.014	0.949	0.650
	(0.005)	(0.005)	(0.004)	(0.000)	(0.000)	(0.001)	(0.003)
IPWE	-0.034	-0.012	-0.077	0.047		0.877	0.710
	(0.009)	(0.009)	(0.009)	(0.003)		(0.003)	(0.003)
AIPWE	-0.032	0.022	-0.041	0.047		0.881	0.705
	(0.008)	(0.009)	(0.008)	(0.003)		(0.002)	(0.003)
	$n = 1000$						
SIM	-0.013	0.0001	0.000	0.005	0.008	0.961	0.651
	(0.004)	(0.004)	(0.003)	(0.000)	(0.000)	(0.001)	(0.002)
IPWE	-0.023	-0.010	-0.053	0.031		0.900	0.690
	(0.007)	(0.007)	(0.007)	(0.002)		(0.002)	(0.002)
AIPWE	-0.021	0.016	-0.021	0.031		0.907	0.688
	(0.006)	(0.007)	(0.007)	(0.002)		(0.002)	(0.002)
	$n = 1500$						
SIM	-0.007	-0.001	-0.002	0.003	0.006	0.970	0.647
	(0.003)	(0.003)	(0.002)	(0.000)	(0.000)	(0.001)	(0.002)
IPWE	-0.024	-0.005	-0.031	0.024		0.913	0.677
	(0.007)	(0.006)	(0.005)	(0.002)		(0.002)	(0.002)
AIPWE	-0.029	-0.003	-0.017	0.024		0.920	0.675
	(0.006)	(0.006)	(0.005)	(0.001)		(0.002)	(0.002)

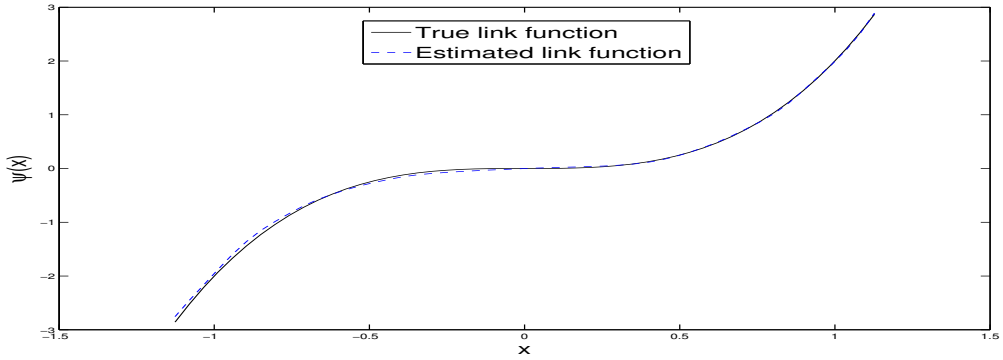
Table 2.3 Estimation and classification results for Example III. Other captions are the same as Table 1.

Method	Bias of $(\beta_1, \beta_2, \beta_3, \beta_4, \beta_5)$					MSE(β)	MSE (ψ)	PCD	Val(0.630)
	$n = 500$								
SIM	-0.003	0.000	-0.003	-0.002	-0.007	0.002	0.011	0.968	0.633
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.000)	(0.000)	(0.001)	(0.003)
IPWE	-0.047	0.055	-0.125	0.041	-0.066	0.058		0.745	0.713
	(0.011)	(0.011)	(0.012)	(0.011)	(0.011)	(0.004)		(0.004)	(0.003)
AIPWE	-0.058	0.055	-0.094	0.035	-0.063	0.058		0.754	0.704
	(0.011)	(0.011)	(0.010)	(0.010)	(0.011)	(0.003)		(0.003)	(0.003)
	$n = 1000$								
SIM	-0.003	0.004	0.003	0.002	0.000	0.001	0.006	0.978	0.635
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.000)	(0.000)	(0.001)	(0.002)
IPWE	-0.036	0.033	-0.093	0.053	-0.038	0.051		0.755	0.688
	(0.009)	(0.010)	(0.011)	(0.010)	(0.010)	(0.003)		(0.003)	(0.002)
AIPWE	-0.061	0.032	-0.078	0.048	-0.014	0.051		0.756	0.682
	(0.009)	(0.009)	(0.010)	(0.010)	(0.008)	(0.003)		(0.003)	(0.002)
	$n = 1500$								
SIM	0.001	-0.001	-0.002	0.000	-0.002	0.001	0.004	0.982	0.632
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.000)	(0.000)	(0.000)	(0.002)
IPWE	-0.039	0.021	-0.071	0.042	-0.052	0.039		0.764	0.675
	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.002)		(0.003)	(0.002)
AIPWE	-0.053	0.052	-0.042	0.029	-0.034	0.039		0.769	0.671
	(0.009)	(0.009)	(0.008)	(0.009)	(0.008)	(0.002)		(0.003)	(0.002)

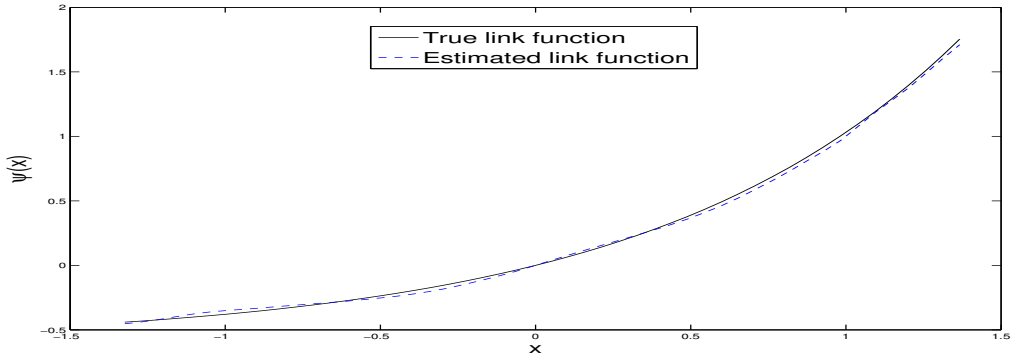
Table 2.4 Inference for the single index parameters of Example 1–3. se1: empirical standard error, se2: mean estimated standard error, cover: empirical coverage rate of 95% confidence intervals.

Example I												
	$n = 500$				$n = 1000$				$n = 1500$			
	bias	se1	se2	cover	bias	se1	se2	cover	bias	se1	se2	cover
β_1	-0.0018	0.030	0.033	0.950	-0.0004	0.020	0.020	0.928	0.0002	0.017	0.017	0.957
β_2	-0.0006	0.028	0.033	0.951	0.0002	0.020	0.020	0.948	0.0006	0.017	0.017	0.952
Example II												
	$n = 500$				$n = 1000$				$n = 1500$			
	bias	se1	se2	cover	bias	se1	se2	cover	bias	se1	se2	cover
β_1	-0.0127	0.104	0.146	0.943	-0.0132	0.079	0.088	0.933	-0.0076	0.063	0.068	0.943
β_2	0.0061	0.102	0.140	0.940	0.0004	0.079	0.087	0.940	-0.0008	0.062	0.067	0.928
β_3	-0.0051	0.079	0.143	0.960	-0.0015	0.064	0.082	0.960	-0.0021	0.050	0.061	0.955
Example III												
	$n = 500$				$n = 1000$				$n = 1500$			
	bias	se1	se2	cover	bias	se1	se2	cover	bias	se1	se2	cover
β_1	-0.003	0.048	0.047	0.953	-0.003	0.031	0.034	0.978	0.000	0.026	0.028	0.975
β_2	0.000	0.047	0.046	0.935	0.004	0.032	0.034	0.958	-0.001	0.027	0.028	0.955
β_3	-0.003	0.044	0.043	0.935	0.003	0.033	0.031	0.923	-0.002	0.025	0.025	0.940
β_4	-0.002	0.043	0.039	0.915	0.002	0.027	0.028	0.958	0.000	0.025	0.023	0.933
β_5	-0.007	0.042	0.039	0.920	0.000	0.030	0.028	0.943	-0.002	0.023	0.023	0.955

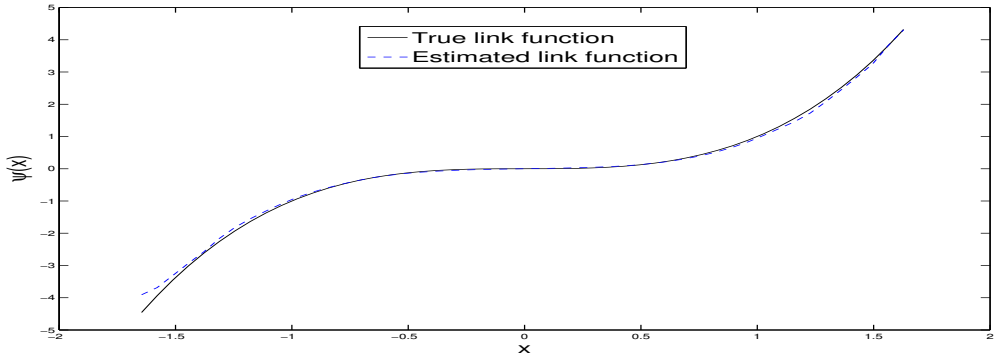
Figure 2.1 Estimation performance for link function based on mean of 10 replications of Example 1–3 when $n = 500$.



Example 1



Example 2



Example 3

Chapter 3

Optimal Treatment Selection with Image Covariates

The personalized medicine has received exponentially increasing attention recently. Different people may respond differently to the same treatment and the same person may respond differently to different treatments as well. Different from the standard treatment discovery framework which is used for finding a single treatment for a homogenous group of patients, personalized medicine involves finding therapies that are tailored to each individual in a heterogeneous group. A treatment regime (strategy) is a function that maps personalized charateristics to available treatment decisions. A key goal in personalized medicine is to find the optimal treatment strategy among all feasible ones, which can provide guidance to physicians and help patients to achieve the most favorable clinical outcome on average.

There is a large collection of literature on statistical methods for estimating the optimal treatment regime based on data from a clinical trial or observational study without imaging covariates. Q-learning (Q denoting "quality") (Watkins and Dayan, 1992; Murphy, 2005; Zhao et al., 2009, 2011; Moodie et al., 2014; Song et al., 2015b) and A-learning (A denoting "advantage") (Murphy, 2003) are two popular regression-type approaches for estimating the optimal treatment regime. The former involves positing parametric models for the regression

outcome on available information and treatment, while the latter requires only part of the regression outcome (representing contrasts among treatments) to be modeled parametrically, along with the propensity scores, the probabilities of observed treatment assignment given patient information. Q-learning can be sensitive to misspecification of the require models, while A-learning enjoys the so-called double robustness property in that the corresponding estimating equations are asymptotically unbiased when either the propensity score or the Q-function (baseline function if we assume the contrast is already correctly specified) is correctly specified. Other than Q- and A-learning, there is another popular class of approaches based on deriving and directly maximizing a consistent estimator of the value function over a specified class of treatment regimes indexed by a finite dimensional parameter. Zhang et al. (2012b) proposed inverse propensity score weighted and augmented inverse propensity score weighted estimators for the value function. Zhao et al. (2012) recast this approach as a weighted classification problem which exploits approximations integrated into classification software to address nonsmooth optimization problem. Song et al. (2015a) also took the treatment regimes as classifier and proposed the penalized outcome weighted learning to deal with high dimensional covariates. They established variable selection consistency property and the asymptotic distribution of related estimator.

While lots of investigators have focused on methods for making a treatment decision based only on scalar covarites at one time point or a sequence of time points referred to as dynamic treatment regimes (Qian and Murphy, 2011; Zhao et al., 2012; Zhang et al., 2012b; Lu et al., 2011; Murphy, 2003; Robins, 2004; Zhao et al., 2009; Laber et al., 2014b), decisions based on both scalars and images have much remained unexplored (McKeague and Qian, 2014; Ciarleglio et al., 2015). McKeague and Qian (2014) proposed methods for estimating and evaluating treatment regimes based on one baseline fucntional covariate. Ciarleglio et al. (2015) proposed an approach to that uses both scalars and multiple functional covariates in the selection of an optimal treatment. However, neither has directly dealt with two or three dimensional imaging features. In this chapter, we develop two different approaches to find the optimal treatment

regime when the contrast function (1.6) is either a linear or nonlinear function of both scalar and imaging covariates.

Formally, we use A to denote treatment assignment taking values of -1 and 1, X to denote the image covariate and/or Z to denote scalar covariate. Let Y be the clinical outcome of interest (assuming large values are desirable). In practice, image covariates are often represented in the form of 2-dimensional matrix or 3-dimensional array. Assume that X is a 2-dimensional matrix of size $N \times N$ which is observed without error and $Z \in R^p$ is a $p \times 1$ vector with the first component being constant one. The decision rule or treatment regime, d, is a function that maps the patient charateristics, (x, z), to the treatment, $a \in \{-1, 1\}$. A patient with covariates $X = x, Z = z$ would receive treatment -1 if $d(x, z) = -1$ and treatment 1 if $d(x, z) = 1$. We aim to find the optimal treatment regime, d^{opt}, which yields the maximum mean clinical outcome.

The first approach assumes that the contrast function is linear, that is,

$$C(X, Z) = \langle X, \beta_0 \rangle + Z^T \theta_0, \quad Y = \psi(X, Z) + AC(X, Z) + \varepsilon \tag{3.1}$$

where $\langle U, V \rangle = \sum_{i,j} u_{i,j} v_{i,j}$ for $U = (u_{i,j}) \in R^{N \times N}$ and $V = (v_{i,j}) \in R^{N \times N}$. Moreover, $\psi(X, Z)$ is the baseline function, $\theta_0 \in R^p$ and $\beta_0(\cdot, \cdot)$ are unknown parameters of interest and $\beta_0(\cdot, \cdot)$ is called the coefficient image/function. ε is the random error with $\mathbb{E}(\varepsilon | X, Z) = 0$ and $\mathbb{E}(\varepsilon^2 | X, Z) = \sigma^2$.

The second does not require the contrast function to be a linear function of both scalar and image covariates. We use convolutional neural networks (CNNs, LeCun et al. (1989)) to directly approximate the contrast function or the Q-function in order to mitigate the risk of mismodeling. CNNs are a family of multi-layer neural networks especially designed to deal with two-dimensional data, such as images and videos. A CNN is a choice of topology or architecture that leverages spatial relationship to reduce the number of parameters which must be learned and thus improves upon back propagation training of similarly-sized general feedforward neural

networks. They are easy to train with GPUs, paired with highly-optimized convolution, while their theoretically-best performance is likely to be only slightly worse. LeCun et al. (1989) first introduced CNNs to exploit the correlation between adjacent pixels in the two-dimensional space and achieved state-of-the-art result for a digit classification problem. Later, researchers have applied CNNs to various machine learning problems with impressive performance, including face detection (Tivive and Bouzerdoum, 2003; Chen et al., 2006), document analysis (Simard et al., 2003), etc. More recently, several papers have shown that CNNs can also achieve outstanding performance on more challenging visual classification tasks (Ciresan et al., 2012; Krizhevsky et al., 2012; Girshick et al., 2014). Most notably, Krizhevsky et al. (2012) demonstrated remarkable performance on the ImageNet 2012 classification benchmark. Their CNN model achieved an error rate of 16.4%, compared to the 2nd place result of 26.1%. The dramatic improvement in performance are due to several factors, including much larger training sets, powerful GPU implementations, making the training of extremely large models practical and better regularization techniques like dropout (Hinton et al., 2012). Zeiler and Fergus (2014) have provided some insight into the impressive performance of CNNs via a novel visualization technique. This motivates us to apply deep learning to find the optimal treatment regime when patients are associated with image covariates.

The rest of chapter is organized as follows. In Section 2, we provide an efficient estimation procedure for model (3.1) and establish the nonasymptotic error bound for the estimation error. In Section 3, we describe convolutional neural networks and employ CNNs to approximate the nonlinear contrast function and Q-function. Extensive simulation studies are presented in Section 4, followed by a discussion section.

3.1 Linear Contrast Function

In this section, we assume that the contrast function is linear in both scalar and image covariates. Let $\Omega = [0, 1)^2$, we know that discrete images are isometric to the space of piecewise-constant

functions defined as

$$\mathcal{I}_N = \left\{ f \in L_2[0,1)^2 : f(s,t) = NX_{jk}, \frac{j-1}{N} \le s < \frac{j}{N}, \frac{k-1}{N} \le t < \frac{k}{N} \right\} \tag{3.2}$$

where X_{jk} is the (j,k)-th pixel value of the image X. We further assume that β_0 is a function of bounded total variation in Ω. The space of bounded total variation in Ω is denoted by $\mathrm{BV}(\Omega)$. Many interesting functions belong to $\mathrm{BV}(\Omega)$. For example, if β_0 belongs to the Sobolev space $W^{1,1}(\mathcal{D})$, i.e., functions with integrable first order partial derivatives. However, the power of total variation in image analysis arises exactly from the relaxation of such constraints. The $\mathrm{BV}(\Omega)$ is much larger than $W^{1,1}(\mathcal{D})$ and contains many interesting piecewise continuous functions with jumps and edges. This is exactly the advantage of using total variation regularization over other familiar regularization methods used in the nonparametric literature.

In the remaining chapter, we treat $\beta = (\beta_{jk}) \in \mathbb{R}^{N\times N}$ as an $N \times N$ block of pixels with β_{jk} as its (j,k) element. Then, we define the discrete total variation of $\beta = (\beta_{jk}) \in \mathbb{R}^{N\times N}$ via its discrete gradient. For any $\beta \in \mathrm{BV}(\Omega)$, the discrete gradient $\nabla : \mathrm{BV}(\Omega) \to \mathbb{R}^{N\times N\times 2}$ is defined by

$$(\nabla\beta)_{jk} = \begin{cases} (\beta_{j+1,k} - \beta_{jk}, \beta_{j,k+1} - \beta_{jk}) & 1 \le j,k \le N-1, \\ (0, \beta_{j,k+1} - \beta_{jk}) & j = N, 1 \le k \le N-1, \\ (\beta_{j+1,k} - \beta_{jk}, 0) & 1 \le j \le N-1, k = N, \\ (0,0) & j = k = N. \end{cases} \tag{3.3}$$

Based on $(\nabla\beta)_{jk} = ((\nabla\beta)_{jk,1}, (\nabla\beta)_{jk,2})^T$, we define both anisotropic and isotropic version of the toal variation norm $\|\beta\|_{TV}$ as follows.

$$\|\beta\|_{TV}^{aniso} = \|\nabla\beta\|_1 = \sum_{jk} \{|(\nabla\beta)_{jk,1}| + |(\nabla\beta)_{jk,2}|\}$$

$$\|\beta\|_{TV}^{iso} = \sum_{jk} |\|(\nabla\beta)_{jk}\|_2 = \sum_{jk} \sqrt{(\nabla\beta)_{jk,1}^2 + (\nabla\beta)_{jk,2}^2}$$

The anisotropic and isotropic induced total variation norms are equivalent up to a factor of $\sqrt{2}$. In this chapter, we will consider the anisotropic case for simplicity and the treatment of the isotropic case is analogous.

We emphasize that there are at least two additional advantages of using bounded variation functions. First, many real images with edges have small total variation since image edges usually reside in a low-dimensional subset of pixels. Second, $\mathrm{BV}(\Omega)$ is mathematically tractable even though it contains many more functions with jumps and edges compared with $W^{1,1}(\mathcal{D})$.

3.1.1 Estimation procedure

In this subsection, we give the detailed estimation procedure of model (3.1). We assume that the propensity score function $\mathbb{P}(A = a|Z, X) = \pi(a|Z, X)$ is known by the trial design. To avoid estimating the nonparametric function $\psi(Z, X)$ in model (3.1), we observe that,

$$\mathbb{E}\left[\frac{AY}{2\pi(A|Z,X)}\middle| Z, X\right] = \langle X, \beta_0\rangle + Z^T\theta_0. \tag{3.4}$$

We propose to solve the following TV minimization:

$$\text{minimize } \|\beta\|_{TV} \quad \text{subject to } \sum_{i=1}^{n}\left[\widetilde{Y}_i - \left(\langle X_i, \beta\rangle + Z_i^T\theta\right)\right]^2 \le \lambda^2, \tag{3.5}$$

where $\widetilde{Y}_i = \frac{A_iY_i}{2\pi(A_i|Z_i,X_i)}$ and λ is a tuning parameter, which controls the noise level.

Let M_X be an $n \times N^2$ design matrix such that the ith row is the vectorized X_i and $Z = (Z_1^T, \ldots, Z_n^T)^T$. We use β to denote both the coefficient matrix and its vectorization. Then we can rewrite (3.5) as the matrix form given by

$$\widehat{\beta} = \arg\min \|\beta\|_{TV} \quad \text{subject to} \quad \|\widetilde{Y} - M_X\beta - Z\theta\|_2 \le \lambda. \tag{3.6}$$

We adapt an algorithm called TVAL3 based on the augmented Lagrangian method (Li, 2011).

Specifically, we solve an equivalent optimization problem given by

$$\min_{w,\beta} \sum_{l=1}^{N^2} \|w_l\|_1 \text{ subject to } \|\widetilde{Y} - M_X\beta - Z\theta\|_2 \leq \lambda \text{ and } D_l\beta = w_l \text{ for all } l,$$

where D_l is an $2 \times N^2$ vector of constants associated with the discrete gradient. Its correponding augmented Lagrangian function is given by

$$L(w, \beta, \theta) = \sum_{l=1}^{N^2} \left\{ \|w_l\|_1 - v_l^T(D_l\beta - w_l) + \frac{\alpha_l}{2}\|D_l\beta - w_l\|_2^2 \right\} + \frac{\gamma}{2}\|\widetilde{Y} - M_X\beta - Z\theta\|_2^2,$$

where v_l, α_l and γ are tuning parameters. We can use either the K-fold cross-validation (CV) or the C_p criterion to select the tuning parameter λ in (3.6). Then the optimization problem is solved iteratively via the following algorithm.

Step 1. Initialize $\beta^{(0)}$.

Step 2. Given $\beta^{(k)}$, we solve for w_l, $l = 1, \ldots, N^2$, by minimizing

$$\|w_l\|_1 - v_l^T(D_l\beta^{(k)} - w_l) + \frac{\alpha_l}{2}\|D_l\beta^{(k)} - w_l\|_2^2.$$

The explicit solution is

$$w_l^{(k)} = \begin{cases} D_l\beta^{(k)} - \frac{v_l+1}{\alpha_l}, & D_l\beta^{(k)} > \frac{v_l+1}{\alpha_l}; \\ 0, & \frac{v_l-1}{\alpha_l} \leq D_l\beta^{(k)} \leq \frac{v_l+1}{\alpha_l}; \\ D_l\beta^{(k)} - \frac{v_l-1}{\alpha_l}, & D_l\beta^{(k)} < \frac{v_l-1}{\alpha_l}. \end{cases} \tag{3.7}$$

Step 3. Given $\left\{ w_l^{(k)} : l = 1, \ldots, N^2 \right\}$, we solve for β, θ by minimizing

$$\sum_{l=1}^{N^2} \left\{ -v_l^T D_l\beta + \frac{\alpha_l}{2}\|D_l\beta - w_l^{(k)}\|_2^2 \right\} + \frac{\gamma}{2}\|\widetilde{Y} - M_X\beta - Z\theta\|_2^2.$$

The explicit solution is

$$\beta^{(k+1)} = \left(\sum_{l=1}^{N^2} \alpha_l D_l^T D_l + \gamma M_X^T (I - P_Z) M_X \right)^{-1} \left(\sum_{l=1}^{N^2} (D_l^T v_l + \alpha_l D_l^T w_l^{(k)}) + \gamma M_X^T (I - P_Z) \widetilde{Y} \right),$$

$$\theta^{(k+1)} = (Z^T Z)^{-1} Z^T (\widetilde{Y} - M_X \beta^{(k+1)}).$$

Step 4. Iterate Step 2 and Step 3 unitil convergence.

3.1.2 Non-asymptotic error bound

In this subsection, we establish the nonasymptotic error bound for the TV estimator $\widehat{\beta}$ based on the model (3.1). We consider two types of distances to measure the error. The first one is a weighted L_2 distance such that

$$\|\widehat{\beta} - \beta_0\|_{X,2} = \left\{ \mathbb{E}(\langle X_{n+1}, \widehat{\beta} - \beta_0 \rangle^2) \right\}^{1/2},$$

where the expectation $\mathbb{E}$ is taken with respect to (Y_{n+1}, X_{n+1}) only. The second one is the TV distance between $\widehat{\beta}$ and β_0, denoted as $\|\widehat{\beta} - \beta_0\|_{TV}$. We derive $\|\widehat{\beta} - \beta_0\|_{X,2}$ and $\|\widehat{\beta} - \beta_0\|_{TV}$ based on Haar wavelet bases. However, Haar wavelets are only for theoretical investigation and we do not estimate the Haar coefficients directly.

Wavelet bases are commonly used to effectively represent images and the Haar wavelet is the simplest possible wavelet. The bivariate Haar wavelet bases for $L_2(\Omega)$ can be constructed as follows. Let $h^0(t) = I_{[0,1)}(t)$ be the indicator function, and the mother wavelet $h^1(t) = 1$ for $t \in [0, 1/2)$ and -1 for $t \in [1/2, 1)$. Starting from the multivariate functions

$$h^d(s,t) = h^{d_1}(s) h^{d_2}(t), \ d \in \{(0,1), (1,0), (1,1)\},$$

the bivariate Haar bases include the indicator function $I_{[0,1)^2}$ and all functions

$$h^d_{j,k}(u,v) = 2^j h^d(2^j x - k),\ d \in \{(0,1),(1,0),(1,1)\},\ x = (u,v),\ j \geq 0,\ k \in \mathbb{Z}^2 \cap 2^j[0,1)^2.$$

The bivariate Haar wavelet bases are orthonormal basis for $L_2[0,1)^2$. Note that discrete images are isometric to the space $\mathcal{I}_N \subset L_2[0,1)^2$ of piecewise constant functions (3.2). Letting $N = 2^J$, the bivariate Haar basis restricted to the N^2 basis functions $\{I_{[0,1)^2}, h^d_{j,k}, j \leq J-1, d \in \{(0,1),(1,0),(1,1)\}, k \in \mathbb{Z}^2 \cap 2^j[0,1)^2\}$ forms an orthonormal basis for $\mathbb{R}^{N\times N}$. Denote by H the discrete bivariate Haar transformation and $\{h_l\}$ the Haar basis, in which $H\beta \in \mathbb{R}^{N\times N}$ contains the bivariate Haar wavelet coefficients of β. Petrushev et al. (1999) established a deep theoretical result on $\mathrm{BV}(\Omega)$ that the Haar wavelet coefficients of $\beta_0 \in \mathrm{BV}(\Omega)$ are in weak ℓ_1. That is, if the Haar coefficients are sorted decreasingly according to their absolute values, then the l-th rearranged coefficient is in absolute value less than $c\|\beta_0\|_{TV}/l$ with c being an absolute constant.

We now introduce the main assumptions of this chapter:

A1. Assume that ε_i are i.i.d sub-Gaussian variables such that $\mathbb{E}(\varepsilon_i|X_i, Z_i) = 0$ and $\mathbb{E}(\varepsilon_i^2|X_i, Z_i) = \sigma^2$. There exists a constant $0 < C_b < \infty$ such that $\psi^2(Z, X) \leq C_b$.

A2. Assume that the coefficient image β_0 is in the space of $N \times N$ blocks of pixel values with bounded total variation, and θ_0 is in a bounded subset of R^p.

A3. Assume that the discrete Haar representation of the image predictor X is $X = \sum_l \rho_l^{1/2} \xi_l h_l$, where ρ_l are positive constants and ξ_l are independently and identically distributed sub-Gaussian random variables with zero mean and unit variance.

A4. For any $\beta \in \mathrm{BV}(\Omega)$, write $\beta = \sum_l \gamma_l h_l$, where the γ_l are the Haar basis coefficients of β. We arrange γ_l in a decreasing order according to their absolute values and denote the sorted coefficients as $\gamma_{(l)}$. Assume that the corresponding sorted $\rho_{(l)}$ associated with the same basis function satisfies $c_1 s^{-2q} \leq \rho_{(s)} \leq c_2 s^{-2q}$ with $q > 1/2$ for each s and two

positive constants c_1, c_2 which may depend on β.

Assumption A1 requires the bounedness of the baseline function, we emphasize that although our current approach can avoid estimating the baseline function, it becomes essentially extra error added to the random error ε, shown in Theorem (3.1.1). Assumption A3 on the wavelet representation of X is reasonable because the discrete wavelet transformation approximately decorrelates data (Vidakovic (2009)). Although we might use Karhunen-Loeve expansion of X, we do not adopt this approach in order to avoid additional complexity associated with the estimation of eigenfunctions. When we sort the Haar wavelets of both β and X, the corresponding basis functions may not follow the same order. Assumption A4 specifies the decay rate of the Haar wavelet coefficients of X. From A3, the predictor images X_i can be written as $X_i = \sum_l \rho_l^{1/2} \xi_{il} h_l$. Let $\widetilde{M}$ be an $n \times N^2$ matrix with the (i,l)-th element being $\xi_{il}/\sqrt{n}$. It is well-known that $\widetilde{M}$ satisfies the restricted isometry property (RIP) with a large probability (Candès et al. (2006a,b)). Specifically, if $n \geq C^{-2} s \log(N^2/s)$, then with probability exceeding $1 - 2\exp(-Cn)$, we have

$$(1-\delta)\|u\|^2 \leq \|\widetilde{M}u\|_2^2 \leq (1+\delta)\|u\|_2^2, \tag{3.8}$$

for all s-sparse vectors $u \in \mathbb{R}^{N^2}$ with a small RIP constant $\delta < C$.

Let $\{\widehat{\gamma}_l\}$ and $\{\gamma_l\}$ be, respectively, the wavelet coefficients of $\widehat{\beta}$ and β_0. It turns out that $\|\widehat{\beta} - \beta_0\|_{X,2} = \{\sum_l \rho_l (\widehat{\gamma}_l - \gamma_l)^2\}^{1/2}$, which is the weighted L_2-norm of the wavelet coefficient difference. On the other hand, since $\|\phi_l\|_{TV} \leq 8$ (Needell and Ward (2013)),

$$\|\widehat{\beta} - \beta_0\|_{TV} \leq \sum_l |\widehat{\gamma}_l - \gamma_l| \|\phi_l\|_{TV} \leq 8 \sum_l |\widehat{\gamma}_l - \gamma_l|, \tag{3.9}$$

which is bounded by the L_1-norm of the wavelet coefficient difference. We obtain the following theorem, whose detailed proof can be found in the Appendix.

Theorem 3.1.1. *Assumptions A1-A4 hold. Let C be an absolute positive constant and $\lambda_0 =$*

$\sqrt{2(C\sigma^2 + C_b)}$. *If* $n \geq Cs^{2q+1}\log(N^2/s^{2q+1})$ *and* $\delta < 1/3$ *in (3.8), then with probability greater than* $1 - 2\exp(-Cn)$*, we have* $\|\widehat{\theta} - \theta_0\|_2 \leq C\lambda_0$*,*

$$\|\widehat{\beta} - \beta_0\|_{X,2} \leq C\left\{\lambda_0 + \frac{1}{(s\log N)^{q+1/2}}\|\nabla\beta_0 - (\nabla\beta_0)_s\|_1\right\}$$

and

$$\|\widehat{\beta} - \beta_0\|_{TV} \leq C\log\left(\frac{N^2}{s}\right)\left\{(s\log N)^{q+1/2}\lambda_0 + \|\nabla\beta_0 - (\nabla\beta_0)_s\|_1\right\}.$$

where $(\nabla\beta_0)_s = \arg\min_{u:s\text{-}sparse}\|\nabla\beta_0 - u\|_1$ *is the best s-sparse approximation to the discrete gradient* $\nabla\beta_0$*.*

3.2 Nonlinear Contrast Function

In this section, we first give a brief introduction to convolutional neural netwoks and then describe the overall architecture of our CNNs. The performance of our CNNs are given in the simulation section.

3.2.1 Convolutional neural networks

Convolutional neural networks have been shown to have excellent performance at tasks such as hand-written digit classification (LeCun et al., 1989), face detection (Tivive and Bouzerdoum, 2003; Chen et al., 2006), document analysis (Simard et al., 2003) and many more recent visual classification tasks (Ciresan et al., 2012; Krizhevsky et al., 2012; Girshick et al., 2014). This motivates us to use CNNs for approximating the contrast function and Q-function to mitigate the risk of mismodeling. However, our problem differs from those image classification problems since we focus on regression and have to deal with extra scalar covariates besides the image covariates.

Convolutional neural networks map an input image, via a series of layers, to a probability

score vector in classification or simply a scalar vector in regression with a target vector. CNNs differ from ordinary neural networks in that CNNs make the explicit assumption that the inputs are images, which allows us to encode certain properties into the architecture via exploiting the correlation structure of image data. In particular, the layers of a CNN have neurons arranged in 3 dimensions: width, height, and depth and the neurons in a layer will only be connected to a small region of the layer before it, instead of all the neurons in a regular fully connected network. This can vastly reduce the number of parameters to be learned.

Specifically, a convolutional neural network usually consists of two types of layers: convolutional layers and pooling layers, as the core building blocks. The parameters of a convolutional layer consist of a set of learnable filters. Each filter is small spatially along width and height, but extends through the full depth of the input volume. We slide each filter across the width and height of the input volume which produces a two dimensional activation map of that filter. As we slide the filter across the input, we are actually computing the dot product between entries of the filter and the input. The full output volume is just the stack of these activation maps for all filters along the depth dimension. As we can see, each entry in the output volume looks at only a small region in the input and shares parameters with neurons in the same activation map since these neurons all result from sliding the same filter (computing dot product using the same filter).

A convolutional layer typically outputs many more activation maps than its input and the adjacent values in an activation map are also highly correlated. Hence, pooling layers were introduced to operate independently on every slice (map) of the input and resize it spatially (along width and height), via the max operation. The most common form is a pooling layer with a filter of size 2×2 applied with a stride of 2 which scales down each slice by a factor of two in both width and height dimensions, discarding 75% of the input neurons. In this case, every max operation would be taking a max over 4 numbers from a 2×2 region in some depth slice. Therefore, the function of pooling layers is to progressively reduce the spatial size of the representation to reduce the amount of parameters and computation in the network, and hence

to also control overfitting.

The final ouput of a convolutional neural network is usually flattened into a vector, followed by one or more fully-connected layers. The architecture of a typical convolutional neural network for classification is shown in figure (3.1). Furthermore, the output of each layer of the entire neural network are passed through a non-linear function, which is often a sigmoid, hyperbolic tangent or rectified linear function ($relu(x) = \max(0, x)$) (Nair and Hinton, 2010). The reason of these nonlinear representation tranformations lies on the fact that a neural network with even one non-linear hidden layer can act as a universal function approximator (Cybenko, 1989) while a neural network, without any non-linear activations, can only act as a linear transformer, regardless of the depth.

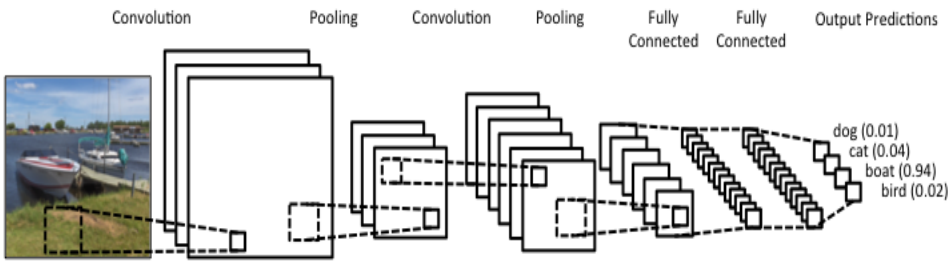


Figure 3.1 Architecture of a convolutional neural network for classification.

3.2.2 Our neural networks

In the simulation, each patient is associated with a image covariate X of dimension 64×64 and a 5 dimensional scalar covariate Z. We introduce two approaches to find the optimal treatment regime, one targeting contrast function with one neural network and the other targeting Q-function with two neural networks (one for each treatment group). In either approach, we pass both scalar and image features separately to two different neural networks, and then the outputs of the two neural networks are concatenated as one vector, followed by two more densely

connected layers. The final output of the entire architecture is a scalar targeting $\frac{AY}{2\pi(A|X,Z)}$ or Y with mean squared loss. In this way, we are directly approximating the contrast function or Q-function via neural networks, since

$$\mathbb{E}\left[\frac{AY}{2\pi(A|Z,X)}\middle|Z,X\right] = C(X,Z), \quad \mathbb{E}\left[Y|Z,X,A\right] = Q(Z,X,A).$$

We can either associate the image feature with a regular neural network after flattening or pass the image feature through a convolutional neural network via a series of convolutional and pooling layers. In the case of a CNN, it has three pairs of convolutional-pooling layers, all of which employ a 3×3 convolution filter and a 2×2 max-pooling window. The choice of 3×3 convolution filter is quite common (LeCun et al., 1989) and more and more architectures tend to use 3×3 filter all through. When bigger filters are really needed, multiple convolutional layers are used instead to reduce the number of parameters and increase nonlinearity. However, we pay the price with increased computational complexity. A 2×2 pooling window is often a good balance between reducing the dimensionality of intermediate representations and keeping adequate feature information. Furthermore, since feature map size decreases with depth (of neural network), layers near the input will tend to have fewer filters while layers higher up can have much more. In practice, we choose the number of filters to make the product of number of feature maps and the number of pixels to be roughly constant across layers. The number of filters directly controls the capacity and so that depends also on the number of available data and the complexity of task. The performance of our neural networks are given in the simulation part.

3.3 Simulation studies

In this section, we conduct extensive Monte Carlo simulations to examine the finite sample performance of our two proposed methods. We simulate X_i from a 64×64 phantom map with $N = 64$ and $4,096$ pixels according to a spatially correlated random process $X_i = \sum_l l^{-q/2}\xi_{il}h_l$

with $q = 0,\ 0.5,$ and 1, where ξ_l are standard normal random variables and the h_l are bivariate Haar wavelet basis functions. We consider four different β_0 images including triangle, oval, T-shape, and checkerboard shapes for the linear case but only T-shape for the nonlinear case. Furthermore, we fix θ as $\theta = (1.0, 1.8, -2.0, 1.6, -2.0)'$, and the first column of Z_i is $\mathbf{1}$ and other columns are generated from multivariate normal with zero mean and compond symmetric covariance structure with correlation 0.3. Random errors are generated from standand normal. Each simulation setting is replicated 100 times.

3.3.1 Linear Case

In the case of linear contrast function, we compare our TV estimators with four competing methods. The first approach (FPCR) is the functional princinpal regression approach (Reiss and Ogden, 2007) by using tensor product cubic B-splines to approximate the coefficient function. The second approach (Matrix-reg) is to estimate β_0 by using a recent development called regularized matrix regression (Zhou and Li, 2014), which treats the coefficient image as a matrix and penalizes the nuclear norm of this matrix. The third one (Lasso) is to calculate the Lasso estimate of β_0. The fourth one (Lasso-Haar) is to calculate the Lasso estimates of the Haar coefficients of β_0 and use the inverse discrete wavelet transform to calculate the estimates of β_0.

We consider the following two examples. One has linear baseline function, the other has nonlinear baseline function.

- Example I : $Y = 0.1(\langle X, \beta\rangle + Z^T\theta) + A\left(\langle X, \beta\rangle + Z^T\theta\right) + \varepsilon,$
- Example II: $Y = \arctan(\langle X, \beta\rangle) + Z_2^2 + 2Z_2Z_3 + A\left(\langle X, \beta\rangle + Z^T\theta\right) + \varepsilon,$

We set $n_1 = 500$ for training set and $n_2 = 200$ for the test set. We repeated each setting 100 times. We present the mean squared error (MSE) of θ and β defined as

$$\text{MSE}(\theta) = \frac{1}{p}\sum_{k=1}^{p}\left(\widehat{\theta}_k - \theta_k\right)^2, \quad \text{MSE}(\beta) = \frac{1}{N^2}\sum_{j=1}^{N}\sum_{k=1}^{N}\left(\widehat{\beta}_{jk} - \beta_{jk}\right)^2$$

We compute the percentage of correct decisions and value function on the $n_2 = 200$ test set as

$$\text{PCD} = 1 - \frac{1}{2n_2}\sum_{k=1}^{n_2}\left|\text{sign}\left(\langle X_k, \widehat{\beta}\rangle + Z_k^T\widehat{\theta}\right) - \text{sign}\left(\langle X_k, \beta_0\rangle + Z_k^T\theta_0\right)\right|,$$

$$\text{VAL} = \frac{1}{n_2}\sum_{k=1}^{n_2}\frac{Y_k\mathbf{1}(A_k = g_k)}{\mathbb{P}(A_k = g_k|Z_k, X_k)}.$$

We also computed the means and standard errors of those MSEs, PCDs and Vals.

Figures 3.2 - 3.5 present the estimated β_0 from a randomly selected training dataset with $q = 0$ and $q = 0.5$, respectively, for the sample size $n = 500$. For all four different shapes, our proposed TV estimates can recover the true shapes and capture the sharp boundaries of the underlying shapes. The functional principal component regression approach uses splines to approximate the predictor images, and it cannot preserve the sharp edges of coefficient estimator for our examples. The matrix regression approach can roughly capture the true shapes when $q = 0$, and unfortunately this method fails for the case when $q = 0.5$, for which the entries of X are spatially correlated. The Lasso method fails for all shapes when $q = 0$ and can barely recover the true shapes when $q = 0.5$. The Lasso estimates of the Haar coefficients can roughly capture the true shapes. However, this method cannot faithfully recover the sharp boundaries of the triangle, oval and T shapes, while it does work very well for the checkerboard shape, since this checkerboard shape is exactly one of the bivariate Haar wavelet basis functions.

Tables 3.2 - 3.9 show that of all five methods, our TV method has significantly smaller mean squared errors for both θ_0 and β_0 except the checkerboard shape with Lasso Haar method since this checkerboard shape is exactly one of the bivariate Haar wavelet basis functions. Moreover, our TV method has larger percentages of correct decisions and value functions for almost all cases.

3.3.2 Nonlinear case

In this subsection, we consider one more example with nonlinear contrast funtions.

- Example III: $Y = \sin\left(\langle X, \beta\rangle + Z^T\theta\right) + 2A\cos\left(\left[\langle X, \beta\rangle + Z^T\theta\right]/3\right) + 0.5\varepsilon$,

- Example IV: $Y = \sin\left(\langle X, \beta\rangle + Z^T\theta\right) + 5A\cos\left(\left[\langle X, \beta\rangle + Z^T\theta\right]/3\right) + 0.5\varepsilon$,

X are generated with $q = 1$ and β is either T-shape or random shape (similarly generated as X). The training of neural networks requires lots of data, however, we set $n_1 = 1000$ for training set and $n_2 = 1000$ for test set in order to be practical in clinical trial design. We consider three different models, all results are shown in table (3.1).

- $(64 \times 64N - 1000N)(5N - 5N) - 500N - 1N$
- $(64 \times 64 - 8C3 - MP2 - 24C3 - MP2 - 72C3 - MP2 - 500N)(5N - 5N) - 250N - 1N$
- TV

64×64 represents the input image, $mC3$ is a 3×3 convolutional layer with m output maps, $MP2$ is a max-pooling layer with 2×2 pooling window size, mN is a fully connected layer with m units. Two adjacent brackets concatenate the outputs of two neural networks. We also run our total variation estimation algorithm, ignoring the fact that the contrast function is nonlinear. It turns out that the TV estimation procedure leads to a PCD of nearly random treatment assignment. All our neural networks give very accurate decisions for patients in the test set and much better clinical outcome on average. Example III shows that when the contrast to baseline ratio is not large enough, targeting Q-functions in the two treatment groups via two convolutional neural networks shows better performance. However, example IV demonstrates the fact that targeting the contrast function is much better when it is large enough compared to the baseline function.

Table 3.1 Example III and IV training size $n_1 = 1000$ and test size $n_2 = 1000$, $q = 1$

Ex	Shape		RegNet(C)	ConvNet(C)	RegNet(Q)	ConvNet(Q)	TV
III	T-Shape	PCD(train)	0.862(0.002)	0.871(0.004)	0.870(0.001)	0.871(0.002)	0.501(0.002)
		PCD(test)	0.774(0.003)	0.830(0.005)	0.812(0.002)	0.850(0.002)	0.504(0.002)
		VAL(train)	1.314(0.006)	1.297(0.010)	1.258(0.006)	1.200(0.009)	1.155(0.007)
		VAL(test)	0.868(0.010)	1.042(0.015)	0.997(0.008)	1.010(0.009)	-0.001(0.007)
	Random-Shape	PCD(train)	0.868(0.002)	0.886(0.003)	0.877(0.001)	0.882(0.002)	0.508(0.002)
		PCD(test)	0.788(0.004)	0.850(0.004)	0.826(0.002)	0.862(0.003)	0.508(0.002)
		VAL(train)	1.323(0.006)	1.324(0.009)	1.272(0.006)	1.228(0.009)	1.279(0.005)
		VAL(test)	0.909(0.012)	1.090(0.013)	1.031(0.007)	1.130(0.008)	0.790(0.008)
IV	T-Shape	PCD(train)	0.916(0.001)	0.918(0.002)	0.891(0.001)	0.881(0.003)	0.502(0.002)
		PCD(test)	0.808(0.003)	0.856(0.003)	0.827(0.002)	0.854(0.002)	0.501(0.002)
		VAL(train)	3.168(0.013)	3.183(0.018)	3.030(0.012)	2.953(0.020)	1.165(0.007)
		VAL(test)	2.420(0.024)	2.771(0.026)	2.573(0.014)	2.769(0.017)	0.009(0.007)
	Random-Shape	PCD(train)	0.921(0.001)	0.927(0.002)	0.899(0.001)	0.894(0.003)	0.513(0.001)
		PCD(test)	0.822(0.003)	0.872(0.003)	0.839(0.002)	0.868(0.003)	0.522(0.002)
		VAL(train)	3.204(0.013)	3.239(0.017)	3.072(0.013)	3.036(0.019)	3.101(0.010)
		VAL(test)	2.523(0.024)	2.872(0.017)	2.664(0.015)	2.864(0.018)	2.199(0.016)

Table 3.2 Example I. MSE(θ) of five methods including TV, FPCR, Matrix-Reg, Lasso, Lasso-Haar for four different shapes, numbers in brackets are the corresponding standard errors of those MSEs.

q	Methods \ Shape	T-shape	Checkerboard	Oval	Triangle
0	TV	0.039 (0.002)	0.102 (0.007)	0.043 (0.003)	0.058 (0.004)
	FPCR	1.742 (0.115)	4.124 (0.269)	1.809 (0.122)	2.261 (0.130)
	Matrix-Reg	0.453 (0.030)	1.257 (0.069)	0.471 (0.032)	0.833 (0.050)
	Lasso	2.016 (0.127)	4.611 (0.303)	2.095 (0.141)	2.407 (0.141)
	Lasso-Haar	0.551 (0.039)	0.050(0.003)	0.373 (0.024)	0.609 (0.048)
0.5	TV	0.012 (0.001)	0.036 (0.003)	0.012 (0.001)	0.016 (0.001)
	FPCR	0.090 (0.006)	0.241 (0.014)	0.091 (0.006)	0.113 (0.007)
	Matrix-Reg	0.070 (0.005)	0.359 (0.021)	0.069 (0.004)	0.098 (0.006)
	Lasso	0.149 (0.010)	0.838 (0.053)	0.199 (0.013)	0.216 (0.015)
	Lasso-Haar	0.044 (0.003)	0.027 (0.002)	0.031 (0.002)	0.050 (0.004)
1	TV	0.008 (0.001)	0.027 (0.002)	0.008 (0.001)	0.010 (0.001)
	FPCR	0.011 (0.001)	0.040 (0.002)	0.012 (0.001)	0.014 (0.001)
	Matrix-Reg	0.032 (0.002)	0.102 (0.006)	0.031 (0.002)	0.047 (0.003)
	Lasso	0.014 (0.001)	0.078 (0.005)	0.015 (0.001)	0.017 (0.001)
	Lasso-Haar	0.011 (0.001)	0.025 (0.002)	0.011 (0.001)	0.015 (0.001)

Table 3.3 Example I. MSE(β) of five methods including TV, FPCR, Matrix-Reg, Lasso, Lasso-Haar for four different shapes, numbers in brackets are the corresponding standard errors of those MSEs.

q	Methods \ Shape	T-shape	Checkerboard	Oval	Triangle
0	TV	0.002 (0.000)	0.007 (0.000)	0.002 (0.000)	0.003 (0.000)
	FPCR	0.177 (0.000)	0.440 (0.000)	0.190 (0.000)	0.217 (0.000)
	Matrix-Reg	0.047 (0.001)	0.132 (0.002)	0.046 (0.000)	0.079 (0.001)
	Lasso	0.202 (0.000)	0.505 (0.001)	0.218 (0.000)	0.250 (0.001)
	Lasso-Haar	0.055 (0.001)	0.000 (0.000)	0.036 (0.000)	0.056 (0.000)
0.5	TV	0.009 (0.000)	0.023 (0.000)	0.009 (0.000)	0.013 (0.000)
	FPCR	0.117 (0.000)	0.215 (0.001)	0.126 (0.000)	0.132 (0.000)
	Matrix-Reg	0.092 (0.000)	0.263 (0.000)	0.086 (0.000)	0.101 (0.000)
	Lasso	0.334 (0.001)	1.431 (0.006)	0.381 (0.001)	0.467 (0.002)
	Lasso-Haar	0.076 (0.001)	0.003 (0.000)	0.054 (0.001)	0.080 (0.001)
1	TV	0.038 (0.001)	0.041 (0.001)	0.024 (0.000)	0.035 (0.000)
	FPCR	0.195 (0.001)	0.493 (0.003)	0.209 (0.001)	0.235 (0.001)
	Matrix-Reg	0.124 (0.000)	0.256 (0.000)	0.137 (0.000)	0.137 (0.000)
	Lasso	0.683 (0.004)	3.332 (0.012)	0.758 (0.004)	0.958 (0.005)
	Lasso-Haar	0.298 (0.010)	0.118 (0.013)	0.271 (0.008)	0.412 (0.014)

Table 3.4 Example I. PCD of five methods including TV, FPCR, Matrix-Reg, Lasso, Lasso-Haar for four different shapes, numbers in brackets are the corresponding standard errors of those PCDs.

q	Methods \ Shape	T-shape	Checkerboard	Oval	Triangle
0	TV	0.966 (0.001)	0.960 (0.002)	0.968 (0.001)	0.964 (0.001)
	FPCR	0.618 (0.004)	0.612 (0.003)	0.617 (0.003)	0.618 (0.004)
	Matrix-Reg	0.873 (0.003)	0.859 (0.003)	0.871 (0.003)	0.829 (0.003)
	Lasso	0.538 (0.004)	0.527 (0.004)	0.541 (0.003)	0.540 (0.004)
	Lasso-Haar	0.828 (0.003)	0.993 (0.001)	0.871 (0.002)	0.846 (0.003)
0.5	TV	0.975 (0.001)	0.979 (0.001)	0.976 (0.001)	0.978 (0.001)
	FPCR	0.902 (0.002)	0.924 (0.002)	0.899 (0.002)	0.914 (0.002)
	Matrix-Reg	0.912 (0.002)	0.889 (0.002)	0.920 (0.002)	0.917 (0.002)
	Lasso	0.848 (0.003)	0.817 (0.003)	0.839 (0.003)	0.850 (0.003)
	Lasso-Haar	0.931 (0.002)	0.993 (0.001)	0.945 (0.002)	0.937 (0.002)
1	TV	0.983 (0.001)	0.989 (0.001)	0.986 (0.001)	0.987 (0.001)
	FPCR	0.969 (0.001)	0.975 (0.001)	0.967 (0.001)	0.969 (0.001)
	Matrix-Reg	0.941 (0.002)	0.959 (0.001)	0.951 (0.001)	0.948 (0.002)
	Lasso	0.961 (0.002)	0.955 (0.002)	0.961 (0.001)	0.962 (0.001)
	Lasso-Haar	0.969 (0.001)	0.993 (0.001)	0.971 (0.001)	0.969 (0.001)

Table 3.5 Example I. Val of five methods including TV, FPCR, Matrix-Reg, Lasso, Lasso-Haar for four different shapes, numbers in brackets are the corresponding standard errors of those Vals.

q	Methods \ Shape	T-shape	Checkerboard	Oval	Triangle
0	TV	22.931 (0.255)	36.033 (0.416)	23.493 (0.266)	25.126 (0.294)
	FPCR	8.248 (0.340)	12.465 (0.472)	8.234 (0.286)	8.802 (0.352)
	Matrix-Reg	21.207 (0.267)	32.890 (0.450)	21.704 (0.278)	21.599 (0.309)
	Lasso	2.963 (0.321)	3.420 (0.479)	2.876 (0.341)	2.870 (0.379)
	Lasso-Haar	19.789 (0.285)	36.276 (0.414)	21.718 (0.280)	22.254 (0.309)
0.5	TV	13.628 (0.147)	27.325 (0.308)	13.988 (0.160)	16.097 (0.182)
	FPCR	13.040 (0.152)	26.596 (0.312)	13.327 (0.165)	15.536 (0.185)
	Matrix-Reg	13.169 (0.152)	25.760 (0.318)	13.569 (0.163)	15.606 (0.181)
	Lasso	12.133 (0.152)	22.968 (0.336)	12.255 (0.168)	14.344 (0.187)
	Lasso-Haar	13.359 (0.150)	27.374 (0.308)	13.810 (0.162)	15.824 (0.182)
1	TV	11.978 (0.129)	26.006 (0.293)	12.500 (0.136)	14.330 (0.160)
	FPCR	11.935 (0.129)	25.940 (0.294)	12.453 (0.136)	14.281 (0.161)
	Matrix-Reg	11.794 (0.129)	25.817 (0.294)	12.368 (0.136)	14.163 (0.162)
	Lasso	11.903 (0.129)	25.756 (0.295)	12.418 (0.136)	14.244 (0.161)
	Lasso-Haar	11.939 (0.128)	26.015 (0.294)	12.468 (0.136)	14.274 (0.160)

Table 3.6 Example II. MSE(θ) of five methods including TV, FPCR, Matrix-Reg, Lasso, Lasso-Haar for four different shapes, numbers in brackets are the corresponding standard errors of those MSEs.

q	Methods \ Shape	T-shape	Checkerboard	Oval	Triangle
0	TV	0.075 (0.006)	0.092 (0.007)	0.080 (0.005)	0.084 (0.005)
	FPCR	1.745 (0.118)	4.140 (0.280)	1.889 (0.128)	2.266 (0.131)
	Matrix-Reg	0.462 (0.030)	1.252 (0.074)	0.519 (0.038)	0.848 (0.052)
	Lasso	2.016 (0.128)	4.636 (0.311)	2.181 (0.149)	2.478 (0.153)
	Lasso-Haar	0.582 (0.047)	0.044 (0.004)	0.418 (0.028)	0.630 (0.045)
0.5	TV	0.055 (0.004)	0.046 (0.003)	0.054 (0.004)	0.052 (0.004)
	FPCR	0.125 (0.010)	0.260 (0.016)	0.133 (0.009)	0.147 (0.009)
	Matrix-Reg	0.104 (0.009)	0.378 (0.022)	0.108 (0.008)	0.132 (0.008)
	Lasso	0.183 (0.012)	0.865 (0.061)	0.230 (0.017)	0.239 (0.016)
	Lasso-Haar	0.088 (0.007)	0.044 (0.003)	0.080 (0.006)	0.097 (0.006)
1	TV	0.045 (0.003)	0.043 (0.003)	0.044 (0.003)	0.043 (0.003)
	FPCR	0.053 (0.004)	0.058 (0.004)	0.055 (0.004)	0.053 (0.004)
	Matrix-Reg	0.065 (0.005)	0.119 (0.007)	0.066 (0.005)	0.078 (0.005)
	Lasso	0.054 (0.004)	0.088 (0.005)	0.056 (0.004)	0.052 (0.004)
	Lasso-Haar	0.052 (0.004)	0.043 (0.003)	0.052 (0.004)	0.054 (0.003)

Table 3.7 Example II. MSE(β) of five methods including TV, FPCR, Matrix-Reg, Lasso, Lasso-Haar for four different shapes, numbers in brackets are the corresponding standard errors of those MSEs.

q	Methods \ Shape	T-shape	Checkerboard	Oval	Triangle
0	TV	0.003 (0.000)	0.005 (0.000)	0.003 (0.000)	0.004 (0.000)
	FPCR	0.178 (0.000)	0.440 (0.000)	0.191 (0.000)	0.218 (0.000)
	Matrix-Reg	0.049 (0.001)	0.129 (0.002)	0.046 (0.000)	0.08 (0.001)
	Lasso	0.202 (0.000)	0.504 (0.001)	0.218 (0.000)	0.249 (0.000)
	Lasso-Haar	0.056 (0.001)	0.000 (0.000)	0.037 (0.000)	0.057 (0.000)
0.5	TV	0.028 (0.001)	0.011 (0.000)	0.025 (0.001)	0.021 (0.001)
	FPCR	0.129 (0.000)	0.217 (0.001)	0.138 (0.000)	0.143 (0.000)
	Matrix-Reg	0.092 (0.000)	0.263 (0.000)	0.087 (0.000)	0.101 (0.000)
	Lasso	0.37 (0.002)	1.433 (0.005)	0.415 (0.002)	0.499 (0.002)
	Lasso-Haar	0.104 (0.001)	0.005 (0.000)	0.076 (0.001)	0.107 (0.001)
1	TV	0.083 (0.002)	0.030 (0.001)	0.050 (0.001)	0.055 (0.001)
	FPCR	0.599 (0.008)	0.607 (0.007)	0.607 (0.007)	0.606 (0.007)
	Matrix-Reg	0.124 (0.000)	0.256 (0.000)	0.137 (0.000)	0.138 (0.000)
	Lasso	1.13 (0.012)	3.389 (0.016)	1.247 (0.015)	1.476 (0.015)
	Lasso-Haar	0.693 (0.027)	0.188 (0.017)	0.565 (0.026)	0.682 (0.035)

Table 3.8 Example II. PCD of five methods including TV, FPCR, Matrix-Reg, Lasso, Lasso-Haar for four different shapes, numbers in brackets are the corresponding standard errors of those PCDs.

q	Methods \ Shape	T-shape	Checkerboard	Oval	Triangle
0	TV	0.961 (0.001)	0.967 (0.002)	0.962 (0.001)	0.960 (0.001)
	FPCR	0.618 (0.004)	0.611 (0.003)	0.617 (0.003)	0.617 (0.004)
	Matrix-Reg	0.869 (0.003)	0.864 (0.003)	0.869 (0.002)	0.826 (0.003)
	Lasso	0.537 (0.004)	0.525 (0.004)	0.540 (0.004)	0.537 (0.004)
	Lasso-Haar	0.827 (0.003)	0.994 (0.001)	0.869 (0.003)	0.845 (0.003)
0.5	TV	0.954 (0.002)	0.985 (0.001)	0.959 (0.001)	0.968 (0.001)
	FPCR	0.895 (0.002)	0.924 (0.002)	0.894 (0.002)	0.907 (0.002)
	Matrix-Reg	0.912 (0.002)	0.889 (0.002)	0.919 (0.002)	0.914 (0.002)
	Lasso	0.844 (0.002)	0.815 (0.003)	0.832 (0.003)	0.843 (0.003)
	Lasso-Haar	0.916 (0.002)	0.993 (0.001)	0.931 (0.002)	0.922 (0.002)
1	TV	0.967 (0.001)	0.989 (0.001)	0.974 (0.001)	0.978 (0.001)
	FPCR	0.946 (0.002)	0.971 (0.001)	0.947 (0.002)	0.952 (0.002)
	Matrix-Reg	0.939 (0.002)	0.959 (0.001)	0.948 (0.001)	0.948 (0.002)
	Lasso	0.945 (0.002)	0.954 (0.001)	0.943 (0.001)	0.948 (0.002)
	Lasso-Haar	0.952 (0.002)	0.991 (0.001)	0.957 (0.001)	0.957 (0.001)

Table 3.9 Example II. Val of five methods including TV, FPCR, Matrix-Reg, Lasso, Lasso-Haar for four different shapes, numbers in brackets are the corresponding standard errors of those Vals.

q	Methods \ Shape	T-shape	Checkerboard	Oval	Triangle
0	TV	24.364 (0.260)	37.588 (0.418)	24.983 (0.278)	26.52 (0.297)
	FPCR	9.665 (0.334)	13.906 (0.455)	9.735 (0.295)	10.239 (0.331)
	Matrix-Reg	22.605 (0.274)	34.590 (0.445)	23.193 (0.286)	22.955 (0.320)
	Lasso	4.263 (0.346)	4.586 (0.495)	4.396 (0.345)	4.100 (0.362)
	Lasso-Haar	21.232 (0.281)	37.779 (0.416)	23.162 (0.288)	23.667 (0.308)
0.5	TV	15.024 (0.16)	28.871 (0.313)	15.385 (0.17)	17.519 (0.19)
	FPCR	14.475 (0.163)	28.139 (0.316)	14.701 (0.174)	16.908 (0.194)
	Matrix-Reg	14.662 (0.161)	27.269 (0.321)	15.019 (0.171)	17.020 (0.188)
	Lasso	13.604 (0.163)	24.478 (0.338)	13.605 (0.175)	15.654 (0.189)
	Lasso-Haar	14.736 (0.161)	28.895 (0.313)	15.156 (0.171)	17.125 (0.189)
1	TV	13.377 (0.138)	27.530 (0.299)	13.955 (0.148)	15.803 (0.168)
	FPCR	13.272 (0.137)	27.430 (0.3)	13.811 (0.147)	15.682 (0.168)
	Matrix-Reg	13.238 (0.137)	27.336 (0.301)	13.842 (0.15)	15.677 (0.171)
	Lasso	13.264 (0.136)	27.289 (0.3)	13.787 (0.149)	15.660 (0.168)
	Lasso-Haar	13.325 (0.137)	27.537 (0.3)	13.890 (0.148)	15.718 (0.169)

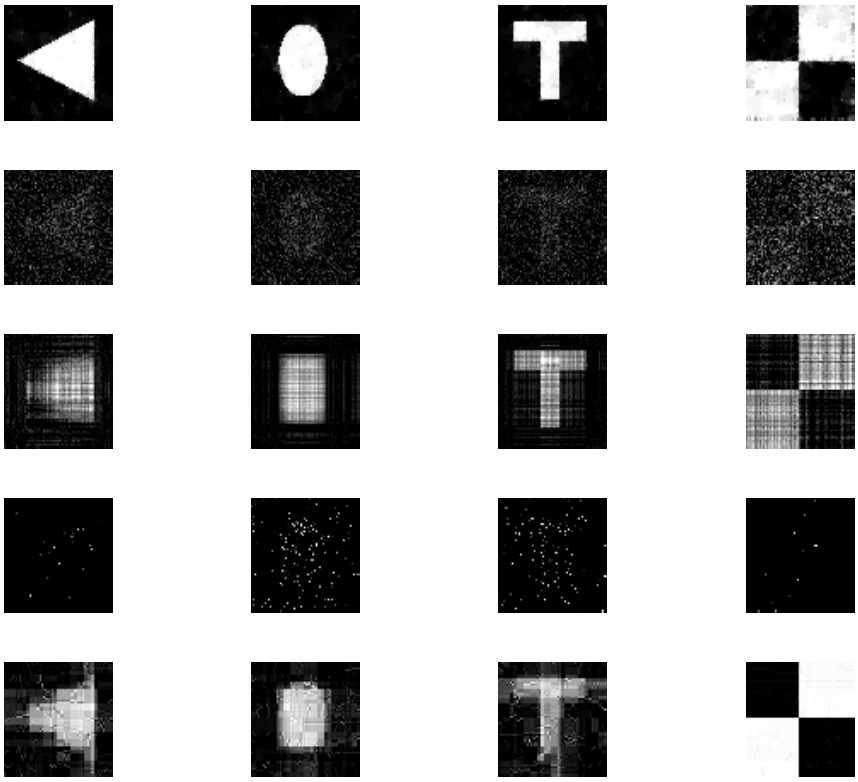


Figure 3.2 The estimated coefficient images from five estimation methods when $q = 0$ and $n = 500$ in Example 1: TV(Top row); FPCR (Second row); Matrix regression (Third Row); Lasso (Fourth row); Lasso-Haar (Fifth row).

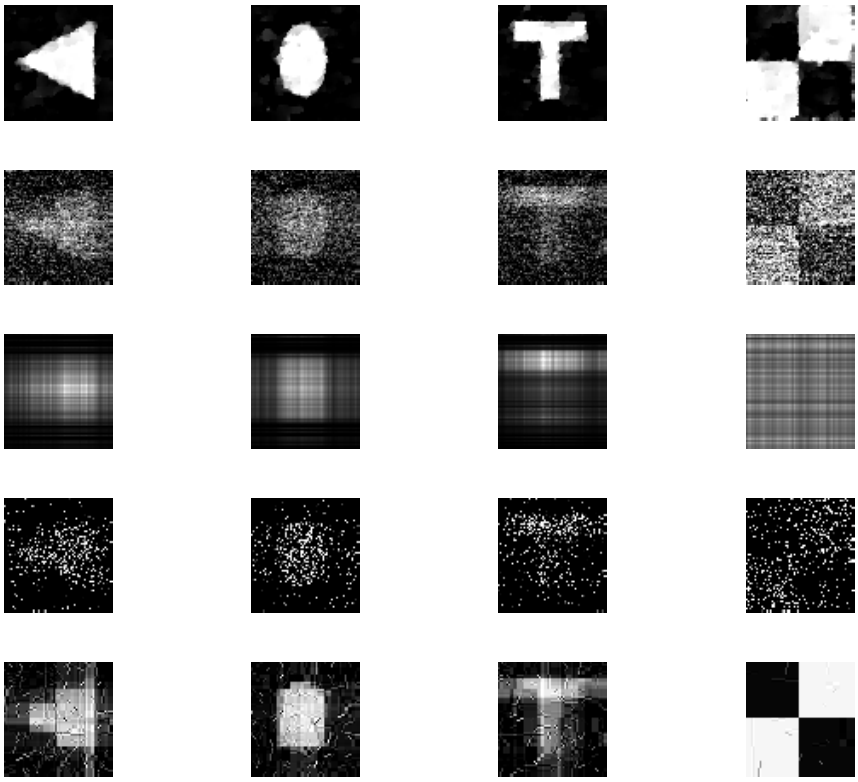


Figure 3.3 The estimated coefficient images from five estimation methods when $q = 0.5$ and $n = 500$ in Example 1: TV(Top row); FPCR (Second row); Matrix regression (Third Row); Lasso (Fourth row); Lasso-Haar (Fifth row).

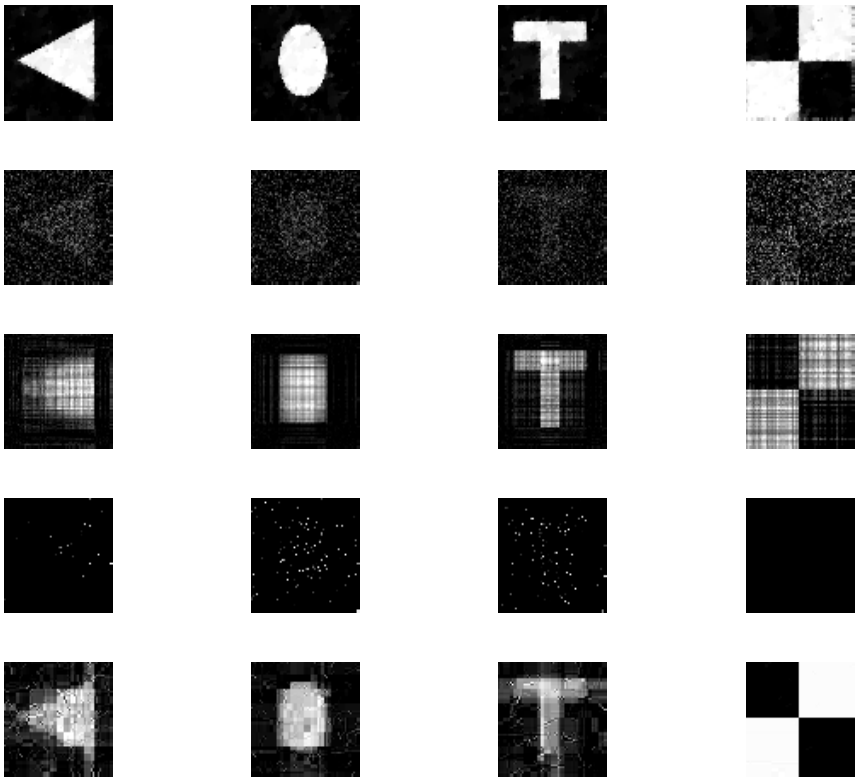


Figure 3.4 The estimated coefficient images from five estimation methods when $q = 0$ and $n = 500$ in Example 2: TV(Top row); FPCR (Second row); Matrix regression (Third Row); Lasso (Fourth row); Lasso-Haar (Fifth row).

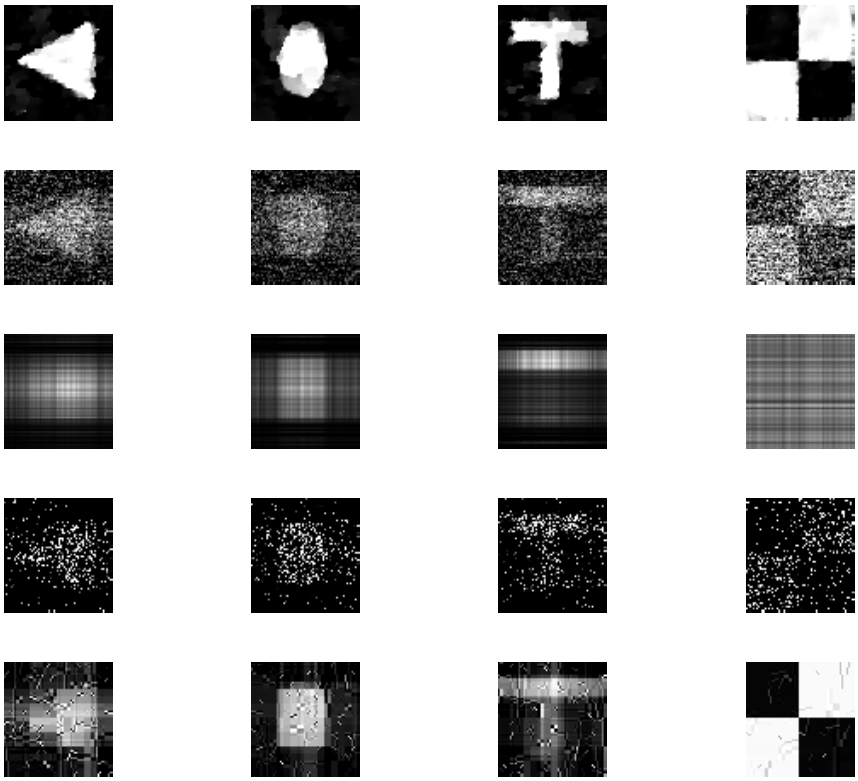


Figure 3.5 The estimated coefficient images from five estimation methods when $q = 0.5$ and $n = 500$ in Example 2: TV(Top row); FPCR (Second row); Matrix regression (Third Row); Lasso (Fourth row); Lasso-Haar (Fifth row).

BIBLIOGRAPHY

Barrett, J. K., Henderson, R., and Rosthøj, S. (2014). Doubly robust estimation of optimal dynamic treatment regimes. *Statistics in biosciences*, 6(2):244–260.

Bather, J. (2000). Decision theory.{A} n introduction to dynamic programming and sequential decisions.

Candès, E. J., Romberg, J., and Tao, T. (2006a). Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *Information Theory, IEEE Transactions on*, 52(2):489–509.

Candès, E. J., Romberg, J. K., and Tao, T. (2006b). Stable signal recovery from incomplete and inaccurate measurements. *Communications on pure and applied mathematics*, 59(8):1207–1223.

Chakraborty, B., Laber, E. B., and Zhao, Y. (2013). Inference for optimal dynamic treatment regimes using an adaptive m-out-of-n bootstrap scheme. *Biometrics*, 69(3):714–723.

Chakraborty, B., Murphy, S., and Strecher, V. (2009). Inference for non-regular parameters in optimal dynamic treatment regimes. *Statistical methods in medical research*.

Chakraborty, B., Murphy, S., and Strecher, V. (2010). Inference for non-regular parameters in optimal dynamic treatment regimes. *Statistical methods in medical research*, 19(3):317–343.

Chen, Y.-N., Han, C.-C., Wang, C.-T., Jeng, B.-S., and Fan, K.-C. (2006). The application of a convolution neural network on face and license plate detection. In *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, volume 3, pages 552–555. IEEE.

Ciarleglio, A., Petkova, E., Ogden, R. T., and Tarpey, T. (2015). Treatment decisions based on scalar and functional baseline covariates. *Biometrics*, 71(4):884–894.

Ciresan, D., Meier, U., and Schmidhuber, J. (2012). Multi-column deep neural networks for image classification. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3642–3649. IEEE.

Craven, M. W. and Shavlik, J. W. (1996). Extracting tree-structured representations of trained networks. *Advances in neural information processing systems*, pages 24–30.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314.

De Boor, C. (1978). A practical guide to splines. *Mathematics of Computation*.

Duan, N. and Li, K.-C. (1991). Slicing regression: a link-free regression method. *The Annals of Statistics*, pages 505–530.

Girshick, R., Donahue, J., Darrell, T., and Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587.

Goldberg, Y., Song, R., Kosorok, M. R., et al. (2013). Adaptive q-learning. In *From Probability to Statistics and Back: High-Dimensional Models and Processes–A Festschrift in Honor of Jon A. Wellner*, pages 150–162. Institute of Mathematical Statistics.

Hardle, W., Hall, P., Ichimura, H., et al. (1993). Optimal smoothing in single-index models. *The annals of Statistics*, 21(1):157–178.

Henderson, R., Ansell, P., and Alshibani, D. (2010). Regret-regression for optimal dynamic treatment regimes. *Biometrics*, 66(4):1192–1201.

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*.

Horowitz, J. L. and Härdle, W. (1996). Direct semiparametric estimation of single-index models with discrete covariates. *Journal of the American Statistical Association*, 91(436):1632–1640.

Hristache, M., Juditsky, A., and Spokoiny, V. (2001). Direct estimation of the index coefficient in a single-index model. *Annals of Statistics*, pages 595–623.

Huang, J. et al. (1996). Efficient estimation for the proportional hazards model with interval censoring. *The Annals of Statistics*, 24(2):540–568.

Ichimura, H. (1993). Semiparametric least squares (sls) and weighted sls estimation of single-index models. *Journal of Econometrics*, 58(1):71–120.

Klein, R. W. and Spady, R. H. (1993). An efficient semiparametric estimator for binary response models. *Econometrica: Journal of the Econometric Society*, pages 387–421.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

Laber, E. B., Linn, K. A., and Stefanski, L. A. (2014a). Interactive model building for q-learning. *Biometrika*, page asu043.

Laber, E. B., Lizotte, D. J., Qian, M., Pelham, W. E., and Murphy, S. A. (2014b). Dynamic treatment regimes: Technical challenges and applications. *Electronic journal of statistics*, 8(1):1225.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551.

Li, C. (2011). *Compressive sensing for 3D data processing tasks: applications, models and algorithms*. PhD thesis, Rice University.

Li, K.-C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316–327.

Li, K.-C. and Duan, N. (1989). Regression analysis under link violation. *The Annals of Statistics*, pages 1009–1052.

Lu, W., Zhang, H. H., and Zeng, D. (2011). Variable selection for optimal treatment decision. *Statistical methods in medical research*, page 0962280211428383.

McKeague, I. W. and Qian, M. (2014). Estimation of treatment policies based on functional predictors. *Statistica Sinica*, 24(3):1461.

Moodie, E. E., Dean, N., and Sun, Y. R. (2014). Q-learning: Flexible learning about useful utilities. *Statistics in Biosciences*, 6(2):223–243.

Moodie, E. E. and Richardson, T. S. (2010). Estimating optimal dynamic regimes: Correcting bias under the null. *Scandinavian Journal of Statistics*, 37(1):126–146.

Moodie, E. E., Richardson, T. S., and Stephens, D. A. (2007). Demystifying optimal dynamic treatment regimes. *Biometrics*, 63(2):447–455.

Murphy, S. A. (2003). Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2):331–355.

Murphy, S. A. (2005). A generalization error for q-learning. *Journal of machine learning research: JMLR*, 6:1073.

Nair, V. and Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 807–814.

Needell, D. and Ward, R. (2013). Stable image reconstruction using total variation minimization. *SIAM Journal on Imaging Sciences*, 6(2):1035–1058.

Petrushev, P., Cohen, A., Xu, H., and DeVore, R. A. (1999). Nonlinear approximation and the space bv (r 2). *American Journal of Mathematics*, 121(3):587–628.

Qian, M. and Murphy, S. A. (2011). Performance guarantees for individualized treatment rules. *Annals of statistics*, 39(2):1180.

Reiss, P. T. and Ogden, R. T. (2007). Functional principal component regression and functional partial least squares. *Journal of the American Statistical Association*, 102(479):984–996.

Robins, J. M. (2004). Optimal structural nested models for optimal sequential decisions. In *Proceedings of the second seattle Symposium in Biostatistics*, pages 189–326. Springer.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of statistics*, pages 34–58.

Schulte, P. J., Tsiatis, A. A., Laber, E. B., and Davidian, M. (2014). Q-and a-learning methods for estimating optimal dynamic treatment regimes. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 29(4):640.

Schumaker, L. (2007). *Spline functions: basic theory*. Cambridge University Press.

Simard, P. Y., Steinkraus, D., and Platt, J. C. (2003). Best practices for convolutional neural networks applied to visual document analysis. In *null*, page 958. IEEE.

Song, R., Kosorok, M., Zeng, D., Zhao, Y., Laber, E., and Yuan, M. (2015a). On sparse representation for optimal individualized treatment selection with penalized outcome weighted learning. *Stat*, 4(1):59–68.

Song, R., Wang, W., Zeng, D., and Kosorok, M. R. (2015b). Penalized q-learning for dynamic treatment regimens. *Statistica Sinica*, 25(3):901–920.

Taylor, J. M., Cheng, W., and Foster, J. C. (2015). Reader reaction to a robust method for estimating optimal treatment regimes by zhang et al.(2012). *Biometrics*, 71(1):267–273.

Thall, P. F., Millikan, R. E., Sung, H.-G., et al. (2000). Evaluating multiple treatment courses in clinical trials. *Statistics in medicine*, 19(8):1011–1028.

Thall, P. F., Sung, H.-G., and Estey, E. H. (2002). Selecting therapeutic strategies based on efficacy and death in multicourse clinical trials. *Journal of the American Statistical Association*, 97(457).

Thall, P. F., Wooten, L. H., Logothetis, C. J., Millikan, R. E., and Tannir, N. M. (2007). Bayesian and frequentist two-stage treatment strategies based on sequential failure times subject to interval censoring. *Statistics in medicine*, 26(26):4687–4702.

Tivive, F. H. C. and Bouzerdoum, A. (2003). A new class of convolutional neural networks (siconnets) and their application of face detection. In *Neural Networks, 2003. Proceedings of the International Joint Conference on*, volume 3, pages 2157–2162. IEEE.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence*. Springer.

Vansteelandt, S., Joffe, M., et al. (2014). Structural nested models and g-estimation: The partially realized promise. *Statistical Science*, 29(4):707–731.

Vidakovic, B. (2009). *Statistical modeling by wavelets*, volume 503. John Wiley & Sons.

Watkins, C. J. and Dayan, P. (1992). Q-learning. *Machine learning*, 8(3-4):279–292.

Xia, Y. (2006). Asymptotic distributions for two estimators of the single-index model. *Econometric Theory*, 22(06):1112–1137.

Zeiler, M. D. and Fergus, R. (2014). Visualizing and understanding convolutional networks. In *Computer vision–ECCV 2014*, pages 818–833. Springer.

Zhang, B., Tsiatis, A. A., Davidian, M., Zhang, M., and Laber, E. (2012a). Estimating optimal treatment regimes from a classification perspective. *Stat*, 1(1):103–114.

Zhang, B., Tsiatis, A. A., Laber, E. B., and Davidian, M. (2012b). A robust method for estimating optimal treatment regimes. *Biometrics*, 68(4):1010–1018.

Zhang, B., Tsiatis, A. A., Laber, E. B., and Davidian, M. (2013). Robust estimation of optimal dynamic treatment regimes for sequential treatment decisions. *Biometrika*, page ast014.

Zhao, Y., Kosorok, M. R., and Zeng, D. (2009). Reinforcement learning design for cancer clinical trials. *Statistics in medicine*, 28(26):3294–3315.

Zhao, Y., Zeng, D., Rush, A. J., and Kosorok, M. R. (2012). Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association*, 107(499):1106–1118.

Zhao, Y., Zeng, D., Socinski, M. A., and Kosorok, M. R. (2011). Reinforcement learning strategies for clinical trials in nonsmall cell lung cancer. *Biometrics*, 67(4):1422–1433.

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

APPENDICES

.1 Supplemental Materials of Chapter 2

Denote $Z = (A, X, Y)$ and $\theta = (\beta, \psi)$. Let P_n denote the empirical measure based on $Z_1 = (A_1, X_1, Y_1)$, ..., $Z_n = (A_n, X_n, Y_n)$ and P denote its expectation. Define

$$\ell(Z;\theta) = \ell(A, X, Y; \beta, \psi) = \left\{ \frac{AY}{2\pi(A|X)} - \psi(\beta^T X) \right\}^2.$$

Then $\widehat{\theta}_n = (\widehat{\beta}_n, \widehat{\psi}_n)$ minimizes $P_n\ell(Z;\theta)$ over $\mathbf{\Theta}_n = \mathcal{B} \times \mathbf{\Psi}_n$ where

$$\mathbf{\Psi}_n = \Big\{ \psi(t) = \sum_{j=1}^{K_n+M} \xi_j N_j(t) : \xi_1 \le \xi_2 \le \cdots \le \xi_{K_n+M}, \quad \sum_{j=1}^{K_n+M} |\xi_j| \le M_n, \ \psi(0) = 0 \Big\}.$$

Finally, let $[a, b]$ be a finite interval containing all $\beta^T x$ for $\beta \in \mathcal{B}$ and x in the support of X.

Proof of Theorem 1.

Since $\widehat{\beta}_n$ is bounded, by choosing a subsequence, we assume $\widehat{\beta}_n$ converges almost surely to a random variable β^*. Clearly $\|\beta^*\| = 1$. Take $\widetilde{\psi}_n$ as the projection of ψ_0 on $\mathbf{\Psi}_n$. According to Schumaker (2007), it satisfies that

$$\|\widetilde{\psi}_n - \psi_0\|_{W^{1,\infty}[a,b]} \le O(K_n^{-k+1})$$

and

$$\|\widetilde{\psi}_n - \psi_0\|_{L^\infty[a,b]} \le O(K_n^{-k}).$$

Recall $W^{1,\infty}$ is the Sobolev norm defined in the space containing all the functions whose derivatives are essentially bounded. Since $\widehat{\theta}_n = (\widehat{\beta}_n, \widehat{\psi}_n)$ is the minimizer of $P_n\ell(Z;\theta)$ over $\mathbf{\Theta}_n = \mathcal{B} \times \mathbf{\Psi}_n$, we have

$$P_n\ell(Z; \widehat{\beta}_n, \widehat{\psi}_n) \le P_n\ell(Z; \beta_0, \widetilde{\psi}_n),$$

we further obtain

$$n^{-1/2}G_n\left\{\ell(Z;\widehat{\beta}_n,\widehat{\psi}_n)-\ell(Z;\beta_0,\widetilde{\psi}_n)\right\}\leq -P\left\{\ell(Z;\widehat{\beta}_n,\widehat{\psi}_n)-\ell(Z;\beta_0,\widetilde{\psi}_n)\right\} \tag{10}$$

where G_n denotes the empirical process $\sqrt{n}(P_n - P)$. We then consider the following class of functions:

$$\mathcal{H}_n=\left\{\psi(\beta^T X):(\beta,\psi)\in\mathbf{\Theta}_n\right\}$$

Since $\beta\in\mathcal{B}$ the unit ball in R^p, we can construct a ϵ-net for $\mathcal{B}$, $\beta_1,\beta_2,\ldots,\beta_K$ with $K=O(1/\epsilon^p)$, such that for any $\beta\in\mathcal{B}$, there is an s such that $|\beta^T X-\beta_s^T X|\leq\epsilon$. Furthermore, for any $(\beta,\psi)\in\mathbf{\Theta}_n$, we have $|\psi'(\beta^T X)|\leq O(M_nK_n)$, so the ϵ-bracket covering number for $\mathcal{H}_n$ is of order $\exp\left\{O(M_nK_n/\epsilon)\right\}/\epsilon^p$ (Corollary 2.7.2, van der Vaart and Wellner (1996)). Consequently, another class of functions, which is defined as

$$\mathcal{F}_n=\left\{\ell(Z;\beta,\psi)-\ell(Z;\beta_0,\widetilde{\psi}_n):(\beta,\psi)\in\mathbf{\Theta}_n\right\}$$

has the bracket covering number of the order

$$N_{[\,]}(\epsilon,\mathcal{F}_n,L_2(P))\leq O\left(\exp\left\{O(M_nK_n/\epsilon)\right\}/\epsilon^p\right).$$

Note that the $L_2(P)$-norm of the envelope function of $\mathcal{F}_n$ is bounded above by $O(M_n^2)$ since $|\psi(\beta^T X)|\leq O(M_n)$. According to Lemma 19.38 of **?**, we obtain that

$$E_P^*\|G_n\|_{\mathcal{F}_n}\lesssim M_n^2\int_0^1\sqrt{\log N_{[\,]}(\epsilon,\mathcal{F}_n,L_2(P))}d\epsilon\leq O(\sqrt{K_nM_n^5}).$$

This implies that the left-hand size of (10) is bounded by $O(\sqrt{K_nM_n^5/n})$. Thus we have

$$P\left\{\ell(Z;\widehat{\beta}_n,\widehat{\psi}_n)-\ell(Z;\beta_0,\widetilde{\psi}_n)\right\}\leq O_p(\sqrt{K_nM_n^5/n}).$$

We further have

$$E\left[\left\{\widehat{\psi}_n(\widehat{\beta}_n^T X) - \psi_0(\beta_0^T X)\right\}^2\right] = P\left\{\ell(Z;\widehat{\beta}_n,\widehat{\psi}_n) - \ell(Z;\beta_0,\psi_0)\right\} \leq O_p(\sqrt{K_n M_n^5/n}) + O(K_n^{-2k}). \quad (11)$$

Note that

$$\begin{aligned} &E\left[\left\{\widehat{\psi}_n(\widehat{\beta}_n^T X) - \psi_0(\beta_0^T X)\right\}^2\right] = \\ &E\left[\left\{\widehat{\psi}_n(\widehat{\beta}_n^T X) - E[\psi_0(\beta_0^T X)|\widehat{\beta}_n^T X]\right\}^2\right] + E\left[\left\{E[\psi_0(\beta_0^T X)|\widehat{\beta}_n^T X] - \psi_0(\beta_0^T X)\right\}^2\right]. \quad (12) \end{aligned}$$

We have

$$E\left[\left\{E[\psi_0(\beta_0^T X)|\widehat{\beta}_n^T X] - \psi_0(\beta_0^T X)\right\}^2\right] \leq O_p(\sqrt{K_n M_n^5/n}) + O(K_n^{-2k}). \quad (13)$$

Following the continuity of $E[\psi_0(\beta_0^T X)|\beta^T X]$ and the dominate convergence theorem, we immediately obtain

$$E[\psi_0(\beta_0^T X)|\beta^{*T} X] = \psi_0(\beta_0^T X).$$

Differentiate both sides with respect to X and evaluate at one point x_0 in its support satisfying $\psi_0'(\beta_0^T x_0) > 0$. Then we conclude that β^* is proportional to β_0. Therefore, $\beta^* = \beta_0$. We reuse inequality (13) and by the mean value and the condition

$$E[\nabla E[\psi_0(\beta_0^T X)|\beta^T X]\Big|_{\beta=\beta_0}^{\otimes 2}] > 0,$$

we thus have

$$\|\widehat{\beta}_n - \beta_0\|^2 \leq O_p(\sqrt{K_n M_n^5/n}) + O(K_n^{-2k}). \quad (14)$$

We reuse (11) and recall $|\widehat{\psi}'| \leq O(K_n M_n)$, it then gives

$$E\left[\left\{\widehat{\psi}_n(\beta_0^T X) - \psi_0(\beta_0^T X)\right\}^2\right] \leq K_n^2 M_n^2 [O_p(\sqrt{K_n M_n^5/n}) + O(K_n^{-2k})].$$

This further gives

$$E\left[\left\{\widehat{\psi}_n(\beta_0^T X) - \widetilde{\psi}_n(\beta_0^T X)\right\}^2\right] \leq K_n^2 M_n^2 [O_p(\sqrt{K_n M_n^5/n}) + O(K_n^{-2k})].$$

Finally, by the fact that the L_2-norm between two functions in $\mathbf{\Psi}_n$ is bounded from below by the Euclidean norm of the corresponding coefficient vectors subject to a constant (De Boor (1978), p. 155), we have

$$\sum_{j=1}^{K_n+M} |\widehat{\xi}_j - \widetilde{\xi}_j|^2 \leq K_n^2 M_n^2 [O_p(\sqrt{K_n M_n^5/n}) + O(K_n^{-2k})].$$

Hence, from the Cauchy-Schwartz inequality, we obtain

$$\sum_{j=1}^{K_n+M} |\widehat{\xi}_j - \widetilde{\xi}_j| \leq \sqrt{K_n^3 M_n^2 [O_p(\sqrt{K_n M_n^5/n}) + O(K_n^{-2k})]}. \tag{15}$$

which indicates that

$$\|\widehat{\psi}_n - \psi_0\|_{L^\infty[a,b]} \leq \sqrt{K_n^3 M_n^2 [O_p(\sqrt{K_n M_n^5/n}) + O(K_n^{-2k})]} + O(K_n^{-k}),$$

which is $o_p(1)$ by the choice of K_n and M_n. The consistency of $\widehat{\psi}_n$ follows. Furthermore,

$$\|\widehat{\psi}_n - \widetilde{\psi}_n\|_{W^{1,\infty}[a,b]} \leq K_n \sum_{j=1}^{K_n+M} |\widehat{\xi}_j - \widetilde{\xi}_j| \leq \sqrt{K_n^5 M_n^2 [O_p(\sqrt{K_n M_n^5/n}) + O(K_n^{-2k})]},$$

which is bounded by the choice of K_n and M_n. It shows that $\widehat{\psi}_n$'s derivative is bounded.

Now we are going to improve the convergence rate of $\widehat{\theta}_n = (\widehat{\beta}_n, \widehat{\psi}_n)$. Since we have shown the consistency of $\widehat{\theta}_n = (\widehat{\beta}_n, \widehat{\psi}_n)$, similar to the proof of convergence rate in Huang et al. (1996),

we may also restrict $\widehat{\psi}_n$ to the following class of functions:

$$\mathbf{\Psi}_n^c = \left\{ \psi(x) = \sum_{j=1}^{K_n+M} \xi_j N_j(x) : \xi_1 \leq \xi_2 \leq \cdots \leq \xi_{K_n+M}, \psi(0) = 0, \|\psi\|_{L^\infty[a,b]} \leq c \right\}$$

where c is a large positive constant.

Let's first re-examine equation (10). For the left-hand side of (10), since

$$\left\{ \beta^T X : \beta \in \mathcal{B} \right\}$$

is a VC-class, by Lemma 2.6.19 of van der Vaart and Wellner (1996),

$$\mathcal{F} \equiv \left\{ \psi(\beta^T X) : \beta \in \mathcal{B}, \|\psi\|_{L^\infty[a,b]} \leq c \right\}$$

is VC-major. By Theorem 2.6.9 of van der Vaart and Wellner (1996), this class has a uniform entropy bounded by

$$\log \sup_Q N(\epsilon, \mathcal{F}, L_2(Q)) \leq c_1 \epsilon^{-c_2},$$

where both c_1 and c_2 are constants. This gives that the left-hand side of (10) is $O_p(n^{-1/2})$. The right hand side of (10) is bounded from above by

$$\begin{aligned} &-P\left[\ell(Z;\widehat{\theta}_n) - \ell(Z;\theta_0)\right] + P\left[\ell(Z;\widetilde{\theta}_n) - \ell(Z;\theta_0)\right] \\ &= -P\left[(\widehat{\psi}_n(\widehat{\beta}_n^T X) - \psi_0({\beta_0}^T X))^2\right] + O(K_n^{-2k}). \end{aligned}$$

Therefore, it gives

$$P\left[(\widehat{\psi}_n(\widehat{\beta}_n^T X) - \psi_0({\beta_0}^T X))^2\right] \leq O_p(n^{-1/2}) + O(K_n^{-2k}). \tag{16}$$

From equation (12), we again have

$$E\left[\left\{E[\psi_0(\beta_0^T X)|\widehat{\beta}_n^T X] - \psi_0(\beta_0^T X)\right\}^2\right] \leq O_p(n^{-1/2}) + O(K_n^{-2k}). \tag{17}$$

By the consistency of $\widehat{\beta}_n$ and the condition

$$E[\nabla E[\psi_0(\beta_0^T X)|\beta^T X]\Big|_{\beta=\beta_0}^{\otimes 2}] > 0,$$

we thus have the improved convergence rate for $\widehat{\beta}_n$:

$$\|\widehat{\beta}_n - \beta_0\|^2 \leq O_p(n^{-1/2}) + O(K_n^{-2k}). \tag{18}$$

Combining with (16) but since now $\widehat{\psi}_n'(x)$ is bounded, we obtain

$$E\left[\left\{\widehat{\psi}_n(\beta_0^T X) - \widetilde{\psi}_n(\beta_0^T X)\right\}^2\right] \leq O_p(n^{-1/2}) + O(K_n^{-2k}) = O_p(n^{-1/2}). \tag{19}$$

We can further improve the rate in (19). To see this, we note that from (18) and (19), in the left-hand side of (10), for a fixed $\nu \in (0, 1/2)$,

$$n^{1/2-\nu}\left[\ell(Z; \widehat{\beta}_n, \widehat{\psi}_n) - \ell(Z; \beta_0, \widetilde{\psi}_n)\right]$$

converges to zero in $L_2(P)$-norm and with probability close to one, it belongs to a class

$$\left\{n^{1/2-\nu}\left[\ell(Z; \beta, \psi) - \ell(Z; \beta_0, \widetilde{\psi}_n)\right] : \right.$$
$$\left. \|\beta - \beta_0\|^2 + \|\psi - \tilde{\psi}_n\|_{L_2(P_X)} \leq Mn^{-1/2}\right\}$$

for a large M. This class satisfies conditions in Theorem 2.11.22 of van der Vaart and Wellner (1996). Thus, the left-hand side of (10) is equal to $o_p(n^{-1+\nu})$. Consequently, we can improve

inequality (18) and (19) to

$$\|\widehat{\beta} - \beta_0\|^2 + \|\widehat{\psi}_n - \psi_0\|^2_{L_2(P_X)} \leq o_p(n^{-1+\nu}) + O(K_n^{-2k}). \tag{20}$$

Correspondingly, inequality (15) can be improved to

$$\sum_{j=1}^{K_n+M} |\widehat{\xi}_j - \widetilde{\xi}_j| \leq \sqrt{K_n\{o_p(n^{-1+\nu}) + O(K_n^{-2k})\}}. \tag{21}$$

This immediately gives

$$\|\widehat{\psi}_n - \widetilde{\psi}_n\|_{W^{1,\infty}[a,b]} \leq K_n\sqrt{K_n\{o_p(n^{-1+\nu}) + O(K_n^{-2k})\}}$$

which gives

$$\|\widehat{\psi}_n - \psi_0\|_{W^{1,\infty}[a,b]} \leq o_p(K_n^{3/2}n^{-1/2+\nu/2}) + O_p(K_n^{-k+3/2}) \tag{22}$$

Theorem 1 then holds if we set $0 < \nu < \min(1 - 3\gamma, 1/2)$.

Proof of Theorem 2

We set $0 < \nu < \min(1/2 - 2\gamma, 1 - 5\gamma)$. First, inequality (21) gives

$$\|\widehat{\psi}_n - \widetilde{\psi}_n\|_{W^{2,\infty}[a,b]} \leq K_n^2\sqrt{K_n\{o_p(n^{-1+\nu}) + O(K_n^{-2k})\}}.$$

Thus, from condition on K_n, $\widehat{\psi}_n''(x)$ is uniformly bounded.

Since $(\widehat{\beta}_n, \widehat{\psi}_n)$ are the minimum argument of $P_n l(A, X, Y; \beta, \psi)$ in $\mathcal{B} \times \mathbf{\Psi}_n$, it holds

$$P_n\left[X\widehat{\psi}_n'(\widehat{\beta}_n^T X)\left\{\frac{AY}{2\pi(A|X)} - \widehat{\psi}_n(\widehat{\beta}_n^T X)\right\}\right] = 0.$$

Thus, we obtain

$$\begin{aligned}&\sqrt{n}(P_n - P)\left[X\widehat{\psi}_n'(\widehat{\beta}_n^T X)\left\{\frac{AY}{2\pi(A|X)} - \widehat{\psi}_n(\widehat{\beta}_n^T X)\right\}\right] \\ &\quad = -\sqrt{n}P\left[X\widehat{\psi}_n'(\widehat{\beta}_n^T X)\left\{\frac{AY}{2\pi(A|X)} - \widehat{\psi}_n(\widehat{\beta}_n^T X)\right\}\right]. \qquad (23)\end{aligned}$$

For the left-hand side of (23), we note that

$$X\widehat{\psi}_n'(\widehat{\beta}_n^T X)\left\{\frac{AY}{2\pi(A|X)} - \widehat{\psi}_n(\widehat{\beta}_n^T X)\right\}$$

belongs to a P-Donsker class because $\left\{\beta^T X : \beta \in \mathcal{B}\right\}$ is a VC class and both $\widehat{\psi}_n$ and $\widehat{\psi}_n'$ are Lipschitz continuous. Moreover, this function converges in $L_2(P)$-norm to

$$X\psi_0'(\beta_0^T X)\left\{\frac{AY}{2\pi(A|X)} - \psi_0(\beta_0^T X)\right\}.$$

Thus, the left-hand side of (23) is equivalent to

$$\sqrt{n}(P_n - P)\left[X\psi_0'(\beta_0^T X)\left\{\frac{AY}{2\pi(A|X)} - \psi_0(\beta_0^T X)\right\}\right] + o_p(1).$$

Note

$$-\sqrt{n}P\left[X\widehat{\psi}_n'(\widehat{\beta}_n^T X)\left\{\frac{AY}{2\pi(A|X)} - \psi_0(\beta_0^T X)\right\}\right] = 0.$$

We further expand the right-hand side of (23) to obtain

$$\sqrt{n}P\left[X\widehat{\psi}_n'(\widehat{\beta}_n^T X)\left\{\widehat{\psi}_n'(\beta_0^T X)X^T(\widehat{\beta}_n - \beta_0) + (\widehat{\psi}_n(\beta_0^T X) - \psi_0(\beta_0^T X))\right\}\right] + O(\sqrt{n}\|\widehat{\beta}_n - \beta_0\|^2)$$

We then replace $X\widehat{\psi}_n'(\widehat{\beta}_n^T X)$ by $X\psi_0'(\beta_0^T X)$. Due to the boundness of $\widehat{\psi}_n''$, we obtain that the

right-hand side of (23) is equivalent to

$$\sqrt{n}P\left[X\psi_0'(\beta_0^TX)\left\{\widehat{\psi}_n'(\beta_0^TX)X^T(\widehat{\beta}_n-\beta_0)+(\widehat{\psi}_n(\beta_0^TX)-\psi_0(\beta_0^TX))\right\}\right]$$
$$+O(\sqrt{n}\|\widehat{\beta}_n-\beta_0\|^2)+O(\sqrt{n}\|\widehat{\psi}_n-\psi_0\|_{L_2(P_X)}^2)+O(\sqrt{n}\|\widehat{\psi}_n'-\psi_0'\|_{L_2(P_X)}^2).$$

From the convergence rates of $(\widehat{\beta}_n, \widehat{\psi}_n, \widehat{\psi}_n')$ in (20) and (22), we finally conclude that the right-hand side of (23) is equal to

$$\sqrt{n}P\Big[X\psi_0'(\beta_0^TX)\left\{\widehat{\psi}_n'(\beta_0^TX)X^T(\widehat{\beta}_n-\beta_0)+(\widehat{\psi}_n(\beta_0^TX)-\psi_0(\beta_0^TX))\right\}\Big]$$
$$+\sqrt{n}o_p(n^{-1+\nu})+\sqrt{n}O(K_n^{-2k})+\sqrt{n}o_p(K_n^2n^{-1+\nu})+\sqrt{n}O(K_n^{-2k+2}).$$

Furthermore, by the choice of K_n, the above expression is equal to

$$\sqrt{n}P\left[\psi_0'(\beta_0^TX)^2XX^T\right](\widehat{\beta}_n-\beta_0)+o_p(1).$$

Combine these results so it holds

$$-\sqrt{n}(P_n-P)\left[X\psi_0'(\beta_0^TX)\left\{\frac{AY}{2\pi(A|X)}-\psi_0(\beta_0^TX)\right\}\right]+o_p(1)$$
$$=\sqrt{n}P\left[\psi_0'(\beta_0^TX)^2XX^T\right](\widehat{\beta}_n-\beta_0).$$

Finally, we note

$$P\left[\psi_0'(\beta_0^TX)^2XX^T\right]=E\left\{\psi_0'(\beta_0^TX)^2E(XX^T|\beta_0^TX)\right\}$$

is non-singular. The asymptotic normality of $\widehat{\beta}_n$ thus follows. Moreover, the asymptotic covariance of $\sqrt{n}(\widehat{\beta}-\beta_0)$ is given by $\Sigma_1^{-1}\Sigma_2\Sigma_1^{-1}$.

.2 Supplemental Materials of Chapter 3

In the following, $a \lesssim b$ means that there exists a universal constant C such that $a \leq Cb$. Without loss of generality, we simply assume $P(A|X,Z) = 0.5$. Take $\lambda = \sqrt{n}\lambda_0$ where $\lambda_0 = \sqrt{2(C\sigma^2 + C_b)}$, then with probability more than $1 - \exp(-Cn)$,

$$\|\widetilde{Y} - M_X\beta_0 - Z\theta_0\|_2 \leq \lambda. \tag{24}$$

In the following argument, we assume that (24) holds. Let $\xi = (\theta, \beta)$, $W = (X, Z)$, and $W_i = (X_i, Z_i), i = 1, \ldots, n$.

$$M_n(\xi) = \sum_{i=1}^{n} \left(\widetilde{Y}_i - \langle X_i, \beta\rangle - Z_i^T\theta\right)^2.$$

Thus the feasibility (24) of (β_0, θ_0) implies that

$$\sum_{i=1}^{n} \left[\langle X_i, \widehat{\beta} - \beta_0\rangle + Z_i^T(\widehat{\theta} - \theta_0)\right]^2 \leq 2(M_n(\widehat{\xi}) + M_n(\xi_0)) \leq 4\lambda^2.$$

Let $g^* = P_{M_X}Z$, then

$$\begin{aligned}
&\sum_{i=1}^{n} \left[\langle X_i, \widehat{\beta} - \beta_0\rangle + Z_i^T(\widehat{\theta} - \theta_0)\right]^2 \\
&= (\widehat{\theta} - \theta_0)^T \left[Z^T(I - P_{M_X})Z\right](\widehat{\theta} - \theta_0) + \sum_{i=1}^{n} \left(\langle X_i, \widehat{\beta} - \beta_0\rangle + (P_{M_X}Z)_i(\widehat{\theta} - \theta_0)\right)^2,
\end{aligned}$$

where P_{M_X} is the projection matrix onto M_X, and $(P_{M_X}Z)_i$ is the ith row of $P_{M_X}Z$. Since $Z^T(I - P_{M_X})Z$ is non-singular, we conclude that $\|\widehat{\theta} - \theta_0\|_2 \lesssim \lambda_0$, and $\|M_X\widehat{\beta} - M_X\beta_0\|_2 \lesssim \lambda_0 n^{1/2}$ follows.

Recall that $\{h_l\}$ is a set of discrete bivariate Haar wavelet basis functions. Write $X_i = \sum_l \xi_{il}\rho_l^{1/2}h_l$, $\beta_0 = \sum_l \gamma_l h_l$, and $\widehat{\beta} = \sum_l \widehat{\gamma}_l h_l$. We aim to derive the error bounds of $\sum_l \rho_l(\widehat{\gamma}_l - \gamma_l)^2$ and $\sum_l |\widehat{\gamma}_l - \gamma_l|$. Denote $\alpha = \widehat{\beta} - \beta_0$ and write $\alpha = \sum_l w_l h_l$, where the $w_l = \widehat{\gamma}_l - \gamma_l$ are the wavelet

coefficients of the difference between the true coefficient image and the estimated coefficient image. We sort w_l in descening order according to their absolute values. Denote the sorted coefficients by $w_{(l)}$. The corresponding ρ_l with the same basis function with h_l is denoted by $\rho_{(l)}$. Note that the $\rho_{(l)}$ are note necessarily sorted, but it is assumed to satisfy the Condition A3.

Let S be the support of the largest s elements of $\nabla\beta_0$. Observe that

$$\|\nabla\widehat{\beta}\|_1 \leq \|\nabla\beta_0\|_1 = \|(\nabla\beta_0)_S\|_1 + \|(\nabla\beta_0)_{S^c}\|_1,$$

and on the other hand,

$$\begin{aligned}\|\nabla\widehat{\beta}\|_1 &= \|(\nabla\alpha)_S + (\nabla\beta_0)_S\|_1 + \|(\nabla\alpha)_{S^c} + (\nabla\beta_0)_{S^c}\|_1 \\ &\geq \|(\nabla\beta_0)_S\|_1 - \|(\nabla\alpha)_S\|_1 - \|(\nabla\beta_0)_{S^c}\|_1 + \|(\nabla\alpha)_{S^c}\|_1.\end{aligned}$$

Combining these two inequalities yields

$$\|(\nabla\alpha)_{S^c}\|_1 \leq \|(\nabla\alpha)_S\|_1 + 2\|\nabla\beta_0 - (\nabla\beta_0)_s\|_1. \tag{25}$$

The cone constraint on the discrete gradient can be transferred to a cone constraint on the wavelet coefficients. Write

$$\alpha = \sum_{j\in\Omega} w_j h_j + \sum_{j\in\Omega^c} w_j h_j,$$

where Ω refer to the set of wavelets that are non-constant over the edges indexed by S, we know that the cardinality of Ω is at most $K = 8s\log N$. Recall that $|w_{(j)}| \leq Cj^{-1}\|\nabla\alpha\|_1$. From (25),

we have

$$\begin{aligned}\sum_{j=K+1}^{N^2}|w_{(j)}| \le \sum_{j=s+1}^{N^2}|w_{(j)}| &\le C\log\left(\frac{N^2}{s}\right)\|\nabla\alpha\|_1 \\ &= C\log\left(\frac{N^2}{s}\right)(\|(\nabla\alpha)_S\|_1 + \|(\nabla\alpha)_{S^c}\|_1) \\ &\le C\log\left(\frac{N^2}{s}\right)(2\|(\nabla\alpha)_S\|_1 + 2\|\nabla\beta_0 - (\nabla\beta_0)_s\|_1) \\ &\lesssim \log\left(\frac{N^2}{s}\right)\left(\sum_{i=1}^{K}|w_{(j)}| + \|\nabla\beta_0 - (\nabla\beta_0)_s\|_1\right) \qquad (26)\end{aligned}$$

where the last inequality holds because $\|\nabla h_j\|_1 \le 8$ (Needell and Ward (2013)), and

$$\|(\nabla\alpha)_S\|_1 = \|\nabla(\sum_{j\in\Omega} w_{(j)}h_{(j)})\|_1 \le \sum_{j\in\Omega}|w_{(j)}|\|\nabla h_{(j)}\|_1 \le 8\sum_{j\in\Omega}|w_{(j)}| \le 8\sum_{i=1}^{K}|w_{(j)}|$$

Furthermore, since $\rho_{(j)}$ is of order j^{-2q} for $q > 0$ from A3, we have

$$\begin{aligned}\sum_{j=K+1}^{N^2}\rho_{(j)}^{1/2}|w_{(j)}| \le C\sum_{j=K+1}^{N^2} j^{-q}j^{-1}\|\nabla\alpha\|_1 &\le \|\nabla\alpha\|_1 \\ &\lesssim \sum_{i=1}^{K}|w_{(j)}| + \|\nabla\beta_0 - (\nabla\beta_0)_s\|_1 \\ &\lesssim K^q\sum_{i=1}^{K}\rho_{(j)}^{1/2}|w_{(j)}| + \|\nabla\beta_0 - (\nabla\beta_0)_s\|_1 \\ &= s^q(\log N)^q\sum_{i=1}^{K}\rho_{(j)}^{1/2}|w_{(j)}| + \|\nabla\beta_0 - (\nabla\beta_0)_s\|_1. \qquad (27)\end{aligned}$$

Now let $\widetilde{S}$ denote the support of s largest entries in the absolute values of α. As shown in Lemma 9 of Needell and Ward (2013), the set $\widetilde{\Omega}$ of wavelets which are non-constant over $\widetilde{S}$ has cardinality at most $K = 8s\log N$. Let

$$\tilde{s} = cs^{2q+1}(\log N)^{2q+1}, \quad d = \lfloor N^2/(4\tilde{s})\rfloor.$$

We may write $\Omega^c = \Omega_1 \cup \Omega_2 \cup \cdots \cup \Omega_d$, where Ω_1 consists of $4\tilde{s}$ largest $|w_{(l)}|$ within Ω^c, Ω_2 consists of next $4\tilde{s}$ largest-magnitude of $|w_{(l)}|$, and so on. Since $\rho_{(l)}$ is of order l^{-2q} and the magnitude of each $\rho_l^{1/2}|w_l|$ in Ω_{j-1} is larger than that in Ω_j up to a constant, we have

$$\left(\sum_{l\in\Omega_j} \rho_l |w_l|^2\right)^{1/2} \lesssim \frac{1}{2\sqrt{\tilde{s}}} \sum_{l\in\Omega_{j-1}} \rho_l^{1/2}|w_l| \text{ for } j = 2, 3, \ldots.$$

Combining this result with (27) yields

$$\begin{aligned}
\sum_{j=2}^{d}\left(\sum_{l\in\Omega_j} \rho_l |w_l|^2\right)^{1/2} &\lesssim \frac{1}{2\sqrt{\tilde{s}}} \sum_{l=K+1}^{N^2} \rho_{(l)}^{1/2}|w_{(l)}| \\
&\lesssim \frac{1}{2\sqrt{\tilde{s}}}\left\{ s^q (\log N)^q \sum_{i=1}^{K} \rho_{(j)}^{1/2}|w_{(j)}| + \|\nabla\beta_0 - (\nabla\beta_0)_s\|_1 \right\} \\
&\lesssim \frac{1}{2\sqrt{K}} \sum_{i=1}^{K} \rho_{(j)}^{1/2}|w_{(j)}| + \frac{1}{2\sqrt{\tilde{s}}}\|\nabla\beta_0 - (\nabla\beta_0)_s\|_1 \\
&\lesssim \frac{1}{2}(\sum_{j=1}^{K} \rho_{(l)}|w_{(l)}|^2)^{1/2} + \frac{1}{2\sqrt{\tilde{s}}}\|\nabla\beta_0 - (\nabla\beta_0)_s\|_1.
\end{aligned}$$

Let $\widetilde{M}$ be an $n \times N^2$ matrix with the (i,l)th element being $n^{-1/2}\xi_{il}$, ρ be a diagonal matrix with the lth diagonal element ρ_l, and γ and w be the wavelet coefficients of β_0 and α, respectively. Therefore, $M_X\beta_0 = \sqrt{n}\widetilde{M}\rho^{1/2}\gamma$ and $M_X\alpha = \sqrt{n}\widetilde{M}\rho^{1/2}w$.

Following the argument in Candès et al. (2006b,a), if $n \geq C^{-2}s\log(N^2/s)$, then $\widetilde{M}$ satisfies the restricted isometry property (RIP) with a large probability: with probability exceeding $1 - 2\exp(-C\delta^2 n)$,

$$(1-\delta)\|u\|^2 \leq \|\widetilde{M}u\|_2^2 \leq (1+\delta)\|u\|_2^2,$$

for all s-sparse vector $u \in R^{N^2}$. Therefore,

$$\sqrt{n}\lambda_0 \gtrsim \sqrt{n}\|\widetilde{M}\rho^{1/2}w\|_2 \geq \sqrt{n}\|\widetilde{M}(\rho^{1/2}w)_{\widetilde{\Omega}} + \widetilde{M}(\rho^{1/2}w)_{\Omega_1}\|_2 - \sqrt{n}\sum_{j=2}^{d}\|\widetilde{M}(\rho^{1/2}w)_{\Omega_j}\|_2$$

$$\geq \sqrt{n(1-\delta)}\|(\rho^{1/2}w)_{\widetilde{\Omega}} + (\rho^{1/2}w)_{\Omega_1}\|_2 - \sqrt{n(1+\delta)}\sum_{j=2}^{d}\|(\rho^{1/2}w)_{\Omega_j}\|_2$$

$$\geq \sqrt{n(1-\delta)}\|(\rho^{1/2}w)_{\widetilde{\Omega}} + (\rho^{1/2}w)_{\Omega_1}\|_2 - \sqrt{n(1+\delta)}\left(\frac{1}{2}\|(\rho^{1/2}w)_{\widetilde{\Omega}}\|_2 + \frac{1}{2\sqrt{\tilde{s}}}\|\nabla\beta_0 - (\nabla\beta_0)_s\|_1\right)$$

$$\geq (\sqrt{1-\delta} - \frac{\sqrt{1+\delta}}{2})\sqrt{n}\|(\rho^{1/2}w)_{\widetilde{\Omega}} + (\rho^{1/2}w)_{\Omega_1}\|_2 - \frac{\sqrt{n(1+\delta)}}{2\sqrt{\tilde{s}}}\|\nabla\beta_0 - (\nabla\beta_0)_s\|_1$$

Since $\delta < 1/3$, we have

$$\sqrt{n}\|(\rho^{1/2}w)_{\widetilde{\Omega}} + (\rho^{1/2}w)_{\Omega_1}\|_2 \lesssim 5\sqrt{n}\lambda_0 + 3\frac{\sqrt{n}}{\sqrt{\tilde{s}}}\|\nabla\beta_0 - (\nabla\beta_0)_s\|_1.$$

$$\|\sum_{j=2}^{d}(\rho^{1/2}w)_{\Omega_j}\|_2 \leq \sum_{j=2}^{d}\|(\rho^{1/2}w)_{\Omega_j}\|_2 \leq \frac{1}{2}\|(\rho^{1/2}w)_{\widetilde{\Omega}} + (\rho^{1/2}w)_{\Omega_1}\|_2 + \frac{\sqrt{n}}{2\sqrt{\tilde{s}}}\|\nabla\beta_0 - (\nabla\beta_0)_s\|_1$$

We arrive at

$$\sqrt{n}\|\rho^{1/2}w\|_2 \lesssim 8\sqrt{n}\lambda_0 + 5\frac{\sqrt{n}}{\sqrt{\tilde{s}}}\|\nabla\beta_0 - (\nabla\beta_0)_s\|_1 \leq \sqrt{n}\lambda_0 + \frac{\sqrt{n}}{(s\log N)^{q+1/2}}\|\nabla\beta_0 - (\nabla\beta_0)_s\|_1.$$

This gives that

$$\|\rho^{1/2}w\|_2 \leq C\left\{\lambda_0 + \frac{1}{(s\log N)^{q+1/2}}\|\nabla\beta_0 - (\nabla\beta_0)_s\|_1\right\}$$

Finally, because

$$\begin{aligned}\sqrt{n}\lambda_0 &\gtrsim \sqrt{n}\|\widetilde{M}\rho^{1/2}w\|_2 \geq \sqrt{n(1-\delta)}\|\rho^{1/2}w\|_2 \\ &\geq \sqrt{n(1-\delta)}\|(\rho^{1/2}w)_{\widetilde{\Omega}}\|_2 \geq \sqrt{n(1-\delta)}K^{-q}\left(\sum_{j=1}^{K}|w_{(j)}|^2\right)^{1/2},\end{aligned}$$

we have

$$\left(\sum_{j=1}^{K}|w_{(j)}|^2\right)^{1/2} \lesssim (s\log N)^q\lambda_0.$$

Combining this with (26) leads to the ℓ_1 error bound since

$$\begin{aligned}\sum_{j=1}^{N^2}|w_{(j)}| &\leq \left(1+\log\left(\frac{N^2}{s}\right)\right)\sum_{j=1}^{K}|w_{(j)}| + \log\left(\frac{N^2}{s}\right)\|\nabla\beta_0 - (\nabla\beta_0)_s\|_1 \\ &\leq \left(1+\log\left(\frac{N^2}{s}\right)\right)K^{1/2}\left(\sum_{j=1}^{K}|w_{(j)}|^2\right)^{1/2} + \log\left(\frac{N^2}{s}\right)\|\nabla\beta_0 - (\nabla\beta_0)_s\|_1 \\ &\lesssim \left(1+\log\left(\frac{N^2}{s}\right)\right)K^{q+1/2}\lambda_0 + \log\left(\frac{N^2}{s}\right)\|\nabla\beta_0 - (\nabla\beta_0)_s\|_1 \\ &\lesssim \log\left(\frac{N^2}{s}\right)\left\{(s\log N)^{q+1/2}\lambda_0 + \|\nabla\beta_0 - (\nabla\beta_0)_s\|_1\right\}.\end{aligned}$$