ABSTRACT

XIAO, WEI. Flexible Methods and Computation for Model Selection and Optimal Treatment Learning. (Under the direction of Wenbin Lu and Hao H. Zhang.)

This dissertation studies the methods and computational algorithm for model selection, and develops robust estimation method for optimal individualized treatment rules. It contains four parts. Chapter 1 gives an overall introduction of the thesis. In Chapter 2, we develop a penalization approach to conduct structure selection for time-varying coefficient Cox models. Specifically, we propose a method which can simultaneously identify the structure of covariate effects in a time-varying coefficient Cox model, i.e., covariates with null effect, constant effect and truly time-varying effect, and estimate the corresponding regression coefficients. Our method is able to identify the underlying true model structure with probability tending to one and estimate the time-varying coefficients consistently. The asymptotic normalities of the resulting estimators for both the constant coefficients and truly time-varying coefficient functions are also established. We demonstrate the performance of our method using simulations and an application to the primary biliary cirrhosis data.

In Chapter 3, we further study and develop the computational algorithm for model/variable selection. We show that the least angle regression (LAR) originally proposed by Efron, Hastie, Johnstone and Tibshirani (2004) for continuous model selection in linear regression can be extended in a fruitful way. In particular, we propose a ConvexLAR algorithm that works for any convex loss function and naturally extends to group selection and data adaptive variable selection. After simple modification it also yields new exact path algorithms for certain penalty methods such as a convex loss function with lasso or group lasso penalty. Variable selection in recurrent event and panel count data analysis, Ada-Boost, and Gaussian graphical model is reconsidered from the ConvexLAR angle.

In Chapter 4, we develop and study new methodologies to personalized medicine. Because different patients may response quite differently to the same drug or treatment, there is increasing interest in developing individualized treatment. In particular, people are eager to find the optimal individualized treatment rule, which is a decision rule that if followed by the whole patient population would lead to the “best” outcome. In this chapter, we propose new estimators based on robust regression with general loss functions to estimate the optimal individualized treatment rules. The new estimators possess the following nice properties: first, they are robust against skewed, heterogeneous, heavy-tailed errors or outliers; second, they are robust against misspecification of the baseline function; third, under certain situations, the new estimator coupled with pinball loss approximately maximizes the outcomes conditional quantile instead of conditional mean, which leads to a different optimal individualized treatment rule from the traditional Q- and A-learning. Consistency and asymptotic normality of the proposed estimators are established. Their empirical performance is demonstrated via extensive simulation studies and an analysis of a AIDS data.

Flexible Methods and Computation for Model Selection and Optimal Treatment Learning

by
Wei Xiao

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

2014

APPROVED BY:

Leonard A. Stefanski	Arnab Maity
Wenbin Lu Co-chair of Advisory Committee	Hao H. Zhang Co-chair of Advisory Committee

DEDICATION

To my parents.

BIOGRAPHY

Wei Xiao was born in Chengdu, Sichuan, China in April 1986. He graduated from Shude High school in 2004 and attended Fudan University afterwards. He received a Bachelor degree in statistics in 2008 from Fudan University. In 2009, he was granted a full scholarship in North Carolina State University to pursue a PhD in statistics. His doctoral dissertation research at North Carolina State University is under the direction of Dr. Wenbin Lu and Dr. Hao Helen Zhang. His dissertation focuses on developing flexible methods and computation for model selection and optimal treatment learning. He will complete the Ph.D. study in May, 2014.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my advisors Dr. Wenbin Lu and Dr. Hao H. Zhang for their enormous support though my graduate study. Their patience, inspiration, encouragement, enthusiasm, and immense knowledge help me in every step of my dissertation research. From them, I learned not only statistical knowledge but also the way to conduct a research. I could not have finished this work without their guidance.

I would also like to extend my appreciation to my committee members, Dr. Leonard A. Stefanski and Dr. Arnab Maity for their advices, and for the wonderful courses they have taught me. I owe special thanks to Dr. Yichao Wu and Dr. Hua Zhou. Specifically, Chapter 3 is based on a joint work with them. Their insight and knowledge are invaluable to me, which bring me to the world of computational statistics. I want also thank Dr. Seth Sullivant, who agrees to serve as my graduate school representative in both my Ph.D. oral preliminary and defense out of his busy schedule.

Thanks to Dr. Lexin Li, Dr. Jason A. Osborne, Dr. Daowen Zhang, Dr. David A. Dickey, Dr. John F. Monahan, Dr. Brian J. Reich, Dr. Min Kang, Dr. Subhashis Ghosal, Dr. Sujit K. Ghosh, Dr. Anastasios Tsiatis, Dr. Eric Laber, Dr. Ana-Maria Staicu and Dr. Dennis Boos for teaching me so many useful courses and giving me advices in my research. Thanks to all my friends at NCSU. The joy we have throughout various parties, sports and travels are unforgettable.

I would also like to thank Quintiles and SAS Institute Inc. for offering me Graduate Industrial Traineeship (GIT) opportunities. Special thanks to Dr. Russell Reeve and Dr. Alex Chien, who are my wonderful supervisors in Quintiles and SAS, respectively. Their support and guidence help me accomplish an excellent work. Their passion about utilizing statistics to solve real world problem always encourages me. I would also like to thank all my former colleagues in Quintiles and SAS. They have made the life of internship so much more colorful.

Last but not least, I would like to express my sincere and wholehearted gratitude to my

parents, for their tremendous love and support. Without their encouragement, I would not be able to achieve everything I have today. They are and always will be my source of strength and courage.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

As technology advances, massive data sets are collected by human beings in all fields, which leads to both promises and challenges for researchers. One important issue in analyzing big volume of data is to find real relationships between the response variable and input covariates in the data, while getting rid of noisy inputs. Variable/model selection as an useful tool to solve this problem has received increasing attention in the past few decades. Another interesting topic is personalized medicine. It is well-known that patients may have heterogeneous responses to the same treatment or drug. Therefore it would optimize the treatment outcome to develop individualized treatment for patients. An individualized treatment rule is a rule that assigns a treatment, from a set of candidate treatments, to a patient according to his/her characteristics, including genetic, physiological, demographic, environmental, and other clinical information, and the optimal individualized treatment rule is one such rule which if assigned to the whole population would lead to the overall "best" outcome. Our main research topic is to find an efficient and robust way to estimate the optimal individualized treatment rule from the data.

1.1 Variable/Model selection

Traditional variable/model selection techniques, such as forward selection, backward elimination, stepwise selection and best-subset selection, are often criticized by their lack of stability, theoretical justification and computational efficiency. In recent years, various modern variable/model selection techniques have been proposed, including but not limited to nonnegative garrote (Breiman, 1995; Yuan and Lin, 2007), LASSO (Tibshirani, 1996, 1997), SCAD (Fan and Li, 2001, 2002), adaptive LASSO (Zou, 2006; Zhang and Lu, 2007), group Lasso (Yuan and Lin, 2006), elastic net (Zou and Hastie, 2005) and MCP (Zhang, 2010). All these methods can be naturally included in the framework of regularization. Most existing methods focus on variable selection in the linear model, and less have been studied for nonparametric models or model structure selection. For example, the identification of linear/nonlinear structure in partially linear regression models or time-invariant/time-varying coefficients in regression models with time varying coefficients.

In Chapter 2, we explore along this line and study the structure selection problem of time-varying coefficient Cox models. The time-varying coefficient Cox model as a natural extension of the standard Cox model is widely used to study the temporal effects of covariates. It relaxes the proportional hazard assumption and can be fitted using either basis expansions (Zucker and Karr, 1990; Yan and Huang, 2012) or local kernel methods (Cai and Sun, 2003; Tian et al., 2005). In a time-varying coefficient Cox model, a critical issue is to correctly identify the model structure, i.e., covariates with null effect, constant effect and truly time-varying effect. By distinguishing covariates with null effect and non-zero effect, we can build a parsimonious model with improved risk prediction. By distinguishing covariates with constant effect and truly time-varying effect, we can build a more interpretable semiparametric model with improved efficiency over a pure nonparametric model. In this chapter, we will develop a double penalization approach to solve this problem, which also simultaneously estimate the corresponding regression coefficients.

The majority of existing work conduct model structure selection for the varying coefficient models by doing hypothesis tests, such as Cai et al. (2000), Huang et al. (2002), Fan and Huang (2005), Kim (2007), Li and Liang (2008) and Wang et al. (2009), etc. Compared with our penalization approach, hypothesis testing based methods are unfavorable due to the lack of theoretical justification and computational efficiency. In addition, it is also extremely difficult to construct a hypothesis test with high power on a large number of covariates simultaneously.

Our proposed methods, KGNG and its variant KGNG2, conduct structure selection and coefficient estimation simultaneously by coupling the kernel-weighted partial likelihood estimation (Cai and Sun, 2003; Tian et al., 2005) with the group nonnegative garrote penalty. The performance of our methods is illustrated using simulations and an application to the primary biliary cirrhosis (PBC) data. Important theoretical properties of the final estimators are established, including selection consistency, estimation consistency and asymptotic normalities of both the constant coefficients and truly time-varying coefficient functions. Based on numerical studies, our method can easily handle high dimensional cases with dimension p over 50. We have developed R code for the proposed methods and make it available at `http://www4.ncsu.edu/~wxiao`.

1.2 Computational Issues for High Dimensional Data

Penalized regression is frequently used in high dimensional data analysis. In penalized regression framework, we usually need to minimize the penalized loss function

$$M(\boldsymbol{\theta}) = L(\boldsymbol{\theta}|\boldsymbol{x}) + \lambda P(\boldsymbol{\theta}) \tag{1.1}$$

with respect to $\boldsymbol{\theta}$, where L is the loss function, P is the penalty and λ is the tuning parameter. The tuning parameter λ controls the magnitude of the penalty, and hence regulates the tradeoffs between goodness of fit and model sparsity. The performance of penalization methods largely depends on the choice of the tuning parameter, and careful selection of the tuning parameter is necessary. We denote the minimizer of (1.1) as $\hat{\boldsymbol{\theta}}(\lambda)$. One important issue in high dimensional data analysis is to develop fast and accurate algorithms to calculate $\hat{\boldsymbol{\theta}}(\lambda)$ over a large number of different λ. The result can then be plugged into tuning criteria such as C_p, AIC and BIC to find the optimal tuning parameter. Typically, there are two ways to conduct the above computation. First, we can construct a grid points of tuning parameter λ and then calculate $\hat{\boldsymbol{\theta}}(\lambda)$ with one value of tuning parameter at a time, using an algorithm such as quadratic programming (QP), the local quadratic approximation (LQA) algorithm (Fan and Li, 2001) or the local linear approximation (LLA) algorithm (Zou and Li, 2008). To improve the tuning result, it is essential to make the grid points of tuning parameters finer and finer, which will on the other hand increase the computation time. This is the main criticism on this type of method. The second way is to apply a solution path algorithm, which can calculate the minimizer $\hat{\boldsymbol{\theta}}(\lambda)$ over all values of λ in one shot. Here, we can apply either the *approximate* path algorithm or the *exact* path algorithm. One important work in this area is the Least Angle Regression (Efron et al., 2004), which is an extremely efficient exact path algorithm for linear regression. The Least Angle Regression (LARS) algorithm results in a piecewise solution path, and is computationally as efficient as Ordinary Least Squares. Rosset and Zhu (2007) derived the sufficient conditions on the loss and penalty functions, which leads to piecewise linear solution paths. They have shown that generalized linear models (GLMs) usually do not have piecewise linear solution paths. Park and Hastie (2006) proposed an approximate path algorithm for GLMs based on the predictor-corrector method of convex optimization. The algorithm's overall accuracy is controlled by a step length parameter. Another approximate path algorithm was proposed in Friedman (2008) for any convex loss with a separable penalty. Wu (2011) proposed an ordinary differential equation based exact path algorithm for GLMs, and extended the algorithm for Cox model considered in Wu (2012). In Chapter 3, we explore along this line and propose a ConvexLAR algorithm that works for any convex loss function and naturally extends to group selection and data adaptive variable selection. The algorithm also leads to a new exact path algorithm for certain penalty methods such as a convex loss function with lasso or group lasso penalty by a simple modification. We apply the algorithm to various contexts including recurrent event and panel count data analysis, Ada-Boost, and Gaussian graphical model. The algorithms are implemented in Matlab and the corresponding code can be downloaded from `http://www4.ncsu.edu/~wxiao`.

1.3 Optimal Treatment Learning

There is increasing interest to discover individual treatment rules to patients due to their heterogenous responses to treatment. Here, an individualized treatment rule is a function that assign a treatment to a patient based on his/her specific characteristics. In particular, we would like to find an optimal individualized treatment rule, which if applied to the whole population would lead to the "best" outcome. For complex diseases such as cancer and AIDS, the optimal individualized treatment rule is usually a dynamic treatment process, which involves a sequence decisions over the time. In this case, we also name it dynamic treatment regimes.

Q-learning (Watkins and Dayan, 1992; Murphy, 2005) and A-learning (Murphy, 2003; Robins, 2004) are two main approaches for estimating optimal dynamic treatment regimes. Q-learning is based on posing a regression model to estimate the conditional expectation of the outcome at each time point, and then applying a backward recursive procedure to fit the model. A-learning, on the other hand, only requires modeling the contrast function of the treatments at each time point, and it is therefore more flexible and robust to a model misspecification. Alternative methods include regret regression (Henderson et al., 2010) and direct value maximization (Zhao et al., 2012; Zhang et al., 2013). Specifically, Zhao et al. (2012) proposed an outcome weighted learning approach using a surrogate loss to approximate a 0-1 loss, and the approach is based on support vector machine framework. Zhang et al. (2013) estimates the dynamic treatment regime by maximizing a doubly robust augmented inverse probability weighted estimator for mean outcome over a restricted class of regimes. The method proposed in Zhang et al. (2013) is robust against model misspecification, but computational intractable when the restricted class of regimes are in high dimension setting.

Variable selection is important to estimate optimal individualized treatment rule in high dimensional data analysis. Related work includes Qian and Murphy (2011) and Lu et al. (2013). The former one employs a flexible model estimation framework to approximate the conditional mean and utilizes ℓ_1 penalty to gain a sparse model. The later one transfers A-learning to a loss-based framework and conducts variable selection with adaptive lasso penalty.

In Chapter 4, we will develop new estimators for optimal individualized treatment rules via robust regression. Compared with standard Q- and A-learning, which both mimic a least squares regression, our new estimators are robust against the skewed, heterogeneous, heavy-tailed errors or outliers. Furthermore, the new estimators are robust against misspecification of the baseline function, which is a nice property shared by A-learning. Last but not least, under certain situations, i.e. when the error term is interacted with the treatment, the new estimator coupled with pinball loss approximately maximizes the outcomes conditional quantile instead of conditional mean, which leads to a different optimal individualized treatment rule from the traditional Q- and A-learning. We demonstrate the performance of the new estimators by simulation studies and an analysis of an AIDS data.

Chapter 2

Joint Structure Selection and Estimation in the Time-varying Coefficient Cox Model

2.1 Introduction

Cox proportional hazards model (Cox, 1972) has become the most popularly used semiparametric model in survival analysis due to its nice hazard interpretation and easy estimation based on partial likelihood principle with elegant counting process-based martingale theory (Andersen and Gill, 1982). However, one main limitation of the standard Cox model is to assume that the hazard ratios stay constant over time, i.e. covariate effects are time-invariant, which may be unrealistic in practical applications. Many alternatives have been proposed to relax the proportional hazards assumption. Among them, the time-varying coefficient Cox model is a natural extension of the standard Cox model by allowing temporal effects of covariates, and has been widely studied in the literature (e.g. Zucker and Karr, 1990; Cai and Sun, 2003; Tian et al., 2005).

An important issue in fitting a time-varying coefficient Cox model is to identify the true model structure, i.e. to distinguish covariates with null effect, constant effect or truly time-varying effect. There are at least two benefits. On one hand, by identifying covariates with null effect and excluding them from the model, we can build a parsimonious model with better risk prediction. On the other hand, by distinguishing covariates with constant effect and truly-time varying effects, we can build a simpler and easier-interpretable semiparametric model comparing with a pure nonparametric model with time-varying coefficients, which will be highly appreciated by empirical investigators since they generally prefer a flexible but easy-interpretable model for data analysis. See Zhang et al. (2002); Fan and Huang (2005); Ahmad et al. (2005); Wang et al. (2009) for more demonstration of the benefits of semiparametric varying-coefficient models comparing with nonparametric varying-coefficient models.

Model selection has been extensively studied in the past few decades. Traditional model selection techniques, such as best-subset selection, coupled with C_p (Mallows, 1973), AIC (Akaike, 1973) and BIC (Schwarz, 1978), separate model selection and model estimation steps and are generally unstable due to the their inherent discreteness (Breiman, 1995) and stochastic errors (Fan and Li, 2001). They also lack of asymptotic selection consistency, which is a desirable asymptotic property to possess. More importantly, they are not computational feasible for data set with moderate to large dimensions as their computation times increase exponentially with the dimension. To overcome these difficulties, various penalization methods have been introduced, for example, nonnegative garrote (Breiman, 1995), LASSO (Tibshirani, 1996, 1997), SCAD (Fan and Li, 2001, 2002) and adaptive LASSO (Zou, 2006; Zhang and Lu, 2007). These methods provide competing performance for simultaneously selecting important variables and estimating their effects. However, most existing penalization methods focus on variable selection for simple linear regression models. Less has been studied for model structure selection, for example, the identification of linear/nonlinear structure in partially linear regression models or time-invariant/time-varying coefficients in regression models with time-varying coefficients. Recently, Zhang et al. (2011) proposed a novel penalization approach in the frame of smoothing spline ANOVA for automatically discovering covariates with null effect, linear effect and nonlinear effect in a partially linear model. For censored data, Yan and Huang (2012) proposed an adaptive group LASSO (AGLASSO) method based on a penalized B-spline approach for model structure selection in a time-varying coefficient Cox model. Specifically, time-varying coefficients are expanded with a set of B-spline basis functions and an adaptive group lasso penalty is used to select between time-independent and time-dependent covariate effects.

In this chapter, we propose an alternative method for automatic model structure selection and coefficient estimation in a time-varying coefficient Cox model by coupling the kernel-weighted partial likelihood estimation (Cai and Sun, 2003; Tian et al., 2005) with the group nonnegative garrote penalty. There are three major motivations for developing this new approach based on local kernel methods. First, in contrast with the spline method proposed in Yan and Huang (2012), our method is able to better capture some local features of time-varying coefficient functions, which are otherwise hard to be captured by the spline method. Second, by using the local kernel estimation, it enables us to rigorously study the asymptotic properties of the proposed estimators for both constant and time-varying coefficients, such as model selection consistency and asymptotic normality, and hence justify the validity of the methods from theoretical perspectives. None of these properties have been established for existing approaches like Yan and Huang (2012). Third, the proposed method provides an automatic and effective way to conduct structure selection for time-varying coefficient Cox model, which can deal with relative large dimension in contrast with all existing methods based on hypothesis testing, such as those studied in Tian et al. (2005) and Liu et al. (2013). The remainder of the chapter is organized as follows. Our proposed kernel group nonnegative garrote (KGNG) method and its variant (KGNG2) are introduced in Section 2.2. Asymptotic properties of KGNG and KGNG2 estimators are presented in Section 2.3. Section 2.4 is devoted to numerical studies, including simulations and an application to the primary biliary cirrhosis data. All the technical proofs are relegated to the Supplement.

2.2 Structure Selection with Kernel Group Nonnegative Garrote

2.2.1 Methods

Consider a random sample of n individuals. Let T_i be the failure time, C_i be the censoring time, and $\boldsymbol{Z}_i$ be a p-vector of covariates for subject i. Conditional on $\boldsymbol{Z}_i$, T_i and C_i are assumed independent. Define $\widetilde{T}_i = \min(T_i, C_i)$ and $\Delta_i = \mathbb{1}(T_i \leq C_i)$. The data consist of the triplets $(\widetilde{T}_i, \boldsymbol{Z}_i, \Delta_i)$, $i = 1, \ldots, n$. The time-varying coefficient Cox model assumes the following form

$$\alpha(t|\boldsymbol{Z}_i) = \alpha_0(t)e^{\boldsymbol{\beta}_0^{\intercal}(t)\boldsymbol{Z}_i}, \tag{2.1}$$

where $\alpha(\cdot|\boldsymbol{Z}_i)$ is the conditional hazard function given covariates, $\alpha_0(\cdot)$ is a completely unspecified baseline hazard function, and $\boldsymbol{\beta}_0(t) = (\beta_{01}(t), \ldots, \beta_{0p}(t))^{\intercal}$ is a p-dimensional smooth function of t.

Without loss of generality, we assume $\boldsymbol{\beta}_0(t) = (\boldsymbol{\beta}_O^{\intercal}(t), \boldsymbol{\beta}_C^{\intercal}(t), \boldsymbol{\beta}_{NC}^{\intercal}(t))^{\intercal}$, where $\boldsymbol{\beta}_O(t) \in \mathbb{R}^{p_1}$, $\boldsymbol{\beta}_C(t) \in \mathbb{R}^{p_2}$ and $\boldsymbol{\beta}_{NC}(t) \in \mathbb{R}^{p_3}$ correspond to covariates with null effect, constant effect and truly time-varying effect, respectively, and $p = p_1 + p_2 + p_3$. Denote the corresponding index sets of the above three classes by I_O, I_C and I_{NC}, and let $\mathrm{I} = \{1, \ldots, p\} = \{\mathrm{I}_O \bigcup \mathrm{I}_C \bigcup \mathrm{I}_{NC}\}$. Our method consists of two steps. In Step 1, for any fixed t, we obtain the initial estimator $\widetilde{\boldsymbol{\beta}}(t) = (\widetilde{\beta}_1(t), \ldots, \widetilde{\beta}_p(t))^{\intercal} \in \mathbb{R}^p$ using the kernel-weighted partial likelihood estimation (Cai and Sun, 2003; Tian et al., 2005), i.e., by maximizing the local partial likelihood

$$L_{1n}(\boldsymbol{\beta}, t) = \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau} K_h(s-t)\left[\boldsymbol{\beta}^{\intercal}\boldsymbol{Z}_i - \log\left(\sum_{j=1}^{n} Y_j(s)e^{\boldsymbol{\beta}^{\intercal}\boldsymbol{Z}_j}\right)\right] dN_i(s), \tag{2.2}$$

with respect to $\boldsymbol{\beta}$, where $K_h(\cdot) = K(\cdot/h)/h$ with $K(\cdot)$ being a symmetric kernel density function and h being the bandwidth parameter, $Y_j(t) = \mathbb{1}(\widetilde{T}_j \geq t)$, $N_i(t) = \mathbb{1}(\widetilde{T}_i \leq t, \Delta_i = 1)$, and τ is a pre-specified constant such that $P(\widetilde{T}_i > \tau) > 0$. Next, we decompose the initial estimator $\widetilde{\boldsymbol{\beta}}(t)$

into the mean part $\widetilde{\boldsymbol{m}} = (\widetilde{m}_1, \ldots, \widetilde{m}_p)^\mathsf{T}$ and the deviation part $\widetilde{\boldsymbol{\beta}}^*(t) = (\widetilde{\beta}_1^*(t), \ldots, \widetilde{\beta}_p^*(t))^\mathsf{T}$, where $\widetilde{m}_k = \tau^{-1}\int_0^\tau \widetilde{\beta}_k(u)du$ and $\widetilde{\beta}_k^*(t) = \widetilde{\beta}_k(t) - \widetilde{m}_k$ for $k = 1, \ldots, p$. Practically, we could choose M grid points $\mathcal{T}_M = \{t_1, \ldots, t_M\}$, equally spaced between 0 and τ, where M is a large positive integer. We then let $\widetilde{m}_k = \sum_{i=1}^M \widetilde{\beta}_k(t_i)/M$ and $\widetilde{\beta}_k^*(t) = \widetilde{\beta}_k(t_j) - \widetilde{m}_k$, where $j = \text{argmin}_k |t_k - t|$ and $t \in [0, \tau]$. In our numerical studies, we set M to be 100, which based on our experiment is large enough to make good approximation of both $\widetilde{m}_k$ and $\widetilde{\beta}_k^*(t)$.

In Step 2, we adapt group nonnegative garrote penalties for structure selection. Denote $\boldsymbol{\lambda}_1 = (\lambda_{11}, \ldots, \lambda_{1p})^\mathsf{T}$, $\boldsymbol{\lambda}_2 = (\lambda_{21}, \ldots, \lambda_{2p})^\mathsf{T}$ as two p-dimensional vectors. We obtain $\widehat{\boldsymbol{\lambda}}_1$ and $\widehat{\boldsymbol{\lambda}}_2$ by minimizing

$$\begin{aligned} Q_{2n}(\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2) = &- \sum_{i=1}^n \int_0^\tau \Bigg[\left(\widetilde{\boldsymbol{m}} \circ \boldsymbol{\lambda}_1 + \widetilde{\boldsymbol{\beta}}^*(s) \circ \boldsymbol{\lambda}_2 \right)^\mathsf{T} \boldsymbol{Z}_i \\ &- \log \left(\sum_{j=1}^n Y_j(s) e^{\left(\widetilde{\boldsymbol{m}} \circ \boldsymbol{\lambda}_1 + \widetilde{\boldsymbol{\beta}}^*(s) \circ \boldsymbol{\lambda}_2 \right)^\mathsf{T} \boldsymbol{Z}_j} \right) \Bigg] dN_i(s) \\ &+ \theta_1 \sum_{j=1}^p \lambda_{1j} + \theta_2 \sum_{j=1}^p \lambda_{2j} \end{aligned} \tag{2.3}$$

subject to $\lambda_{1j} \geq 0$ and $\lambda_{2j} \geq 0$, $j = 1, \cdots, p$, where $\boldsymbol{\theta} = (\theta_1, \theta_2)$ are two-dimensional nonnegative tuning parameters, and $a \circ b$ denotes the Hadamard (element-wise) product of vectors a and b. Then, the proposed KGNG estimator of $\beta_{0k}(t)$ is given by

$$\widehat{\beta}_k(t) = \widehat{\lambda}_{1k}\widetilde{m}_k + \widehat{\lambda}_{2k}\widetilde{\beta}_k^*(t), \; k = 1, \ldots, p, \; t \in [0, \tau]. \tag{2.4}$$

It is interesting to note that the automatic structure selection is achieved by shrinking some components of $\widehat{\boldsymbol{\lambda}}_1$ and $\widehat{\boldsymbol{\lambda}}_2$ to zero. Specifically, we define $\widehat{\mathrm{I}}_O = \{k \in \mathrm{I} : \widehat{\lambda}_{1k} = 0, \widehat{\lambda}_{2k} = 0\}$, $\widehat{\mathrm{I}}_C = \{k \in \mathrm{I} : \widehat{\lambda}_{1k} \neq 0, \widehat{\lambda}_{2k} = 0\}$ and $\widehat{\mathrm{I}}_{NC} = \{k \in \mathrm{I} : \widehat{\lambda}_{2k} \neq 0\}$ as estimated index sets for I_O, I_C and I_{NC}, respectively.

When the number of covariates is large, the inclusion of many noise variables (covariates with null effect) at Step 1 may affect the structure selection performance. Next, we propose a variant of KGNG estimator. Specifically, we add a preliminary step (Step 0) before Steps 1 and 2 to exclude all noise variables prior to structure selection. To do this, we conduct a standard group nonnegative garrote estimation by minimizing

$$Q_{0n}(\boldsymbol{\lambda}^*) = -\sum_{i=1}^{n}\int_0^{\tau}\left[\left(\widetilde{\boldsymbol{\beta}}(s)\circ\boldsymbol{\lambda}^*\right)^{\mathsf{T}}\boldsymbol{Z}_i - \log\left(\sum_{j=1}^{n}Y_j(s)e^{\left(\widetilde{\boldsymbol{\beta}}(s)\circ\boldsymbol{\lambda}^*\right)^{\mathsf{T}}\boldsymbol{Z}_j}\right)\right]dN_i(s)$$
$$+\,\theta^*\sum_{j=1}^{p}\lambda_j^* \tag{2.5}$$

with respect to $\boldsymbol{\lambda}^* = (\lambda_1^*,\ldots,\lambda_p^*)^{\mathsf{T}}$, where $\lambda_j^* \geq 0$, $j = 1,\ldots,p$, and θ^* is a nonnegative tuning parameter. Let $\widehat{\boldsymbol{\lambda}}^* = (\widehat{\lambda}_1^*,\ldots,\widehat{\lambda}_p^*)^{\mathsf{T}}$ denote the resulting minimizer. We exclude the kth covariate if $\widehat{\lambda}_k^* = 0$. Let $\underline{\boldsymbol{Z}}_i$ denote the remaining sub-vector of $\boldsymbol{Z}_i$ by keeping all important covariates. We then implement Steps 1 and 2 with $\boldsymbol{Z}_i$ replaced by $\underline{\boldsymbol{Z}}_i$ for further structure selection. The resulting estimator is denoted as KGNG2.

2.2.2 Computational Aspects

We implemented the proposed method in R, and the corresponding code can be downloaded from the author's web page (`http://www4.ncsu.edu/sim/wxiao/`). In Step 1, $L_{1n}(\boldsymbol{\beta}, t)$ is strictly concave with probability one and thus has a unique solution. The maximization can be realized based on a regular Newton-Raphson iteration or an efficient iterative algorithm proposed in Cai et al. (2000). In Step 2, after proper transformations, the minimization problem of (2.3) is equivalent to finding the lasso solution for a Cox model with time-dependent covariates $\widetilde{\boldsymbol{Z}}_i(s)$ under the nonnegative constraint of regression parameters, where

$$\widetilde{\boldsymbol{Z}}_i(s) = \begin{pmatrix} \widetilde{\boldsymbol{m}}\circ\boldsymbol{Z}_i \\ \widetilde{\boldsymbol{\beta}}^*(s)\circ\boldsymbol{Z}_i \end{pmatrix}.$$

We used R package "penalized" (Goeman, 2010) for this optimization step. The algorithm is based on a combination of gradient ascent optimization and Newton-Raphson algorithm, which can also incorporate nonnegative constraints on the parameters (Goeman, 2010). Moreover, the minimization in the preliminary Step 0 is also equivalent to finding a lasso solution for a Cox model with time-dependent covariates $\widetilde{\boldsymbol{\beta}}(s) \circ \boldsymbol{Z}_i$. Therefore it can be computed similarly with existing R packages.

2.2.3 Tuning Procedure

For computing KGNG, two sets of tuning parameters need to be chosen properly, i.e., the bandwidth h at the maximum local partial likelihood estimation step (2.2) and (θ_1, θ_2) at the group nonnegative garrote estimation step (2.3). To choose h, we use a K-fold cross-validation method as suggested in Tian et al. (2005). First, we randomly split the data set into K roughly equal-sized parts. Then, for each $k = 1, \ldots, K$, we delete the kth part and fit the time-varying coefficient Cox model with the other $K-1$ parts. Next, we compute the prediction error $PE_k(h)$, which measures how well the fitted model predicts the kth part of the data. Here,

$$PE_k(h) = -\sum_{i \in \mathrm{I}_k} \int_0^\tau \left[\widehat{\boldsymbol{\beta}}^{(-k)}(s)^\intercal \boldsymbol{Z}_i - \log\left(\sum_{j \in \mathrm{I}_k} Y_j(s) e^{\widehat{\boldsymbol{\beta}}^{(-k)}(s)^\intercal \boldsymbol{Z}_j} \right) \right] dN_i(s),$$

where I_k is the index set for the kth part of the data and $\widehat{\boldsymbol{\beta}}^{(-k)}(t)$ is the maximum local partial likelihood estimator calculated with the kth part of the data deleted. Last, the optimal h is obtained by minimizing the total prediction error $PE(h) = \sum_{k=1}^K PE_k(h)$.

For (θ_1, θ_2), we consider a set of bivariate grid values, and choose the optimal (θ_1, θ_2) by minimizing the following BIC criterion

$$\textbf{BIC} = -2\log(\textbf{partial likelihood})/n + \log n/n \times df_1 + \log(nh^*)/(nh^*) \times df_2,$$

where df_1 and df_2 are the number of nonzero components in $\widehat{\boldsymbol{\lambda}}_1$ and $\widehat{\boldsymbol{\lambda}}_2$, respectively, and $h^* = h/\tau$ is the effective bandwidth when we scale τ to 1. Note that the effective sample size nh^* is used here for the time-varying components instead of the original n to account for the fact that $\beta_i^*(t)$ is estimated locally. Similar strategy is adopted in Wang and Xia (2009) and Hu and Xia (2012).

Similarly, θ^* in the preliminary Step 0 can be chosen by minimizing the BIC criterion

$$\mathbf{BIC} = -2\log(\mathbf{partial\ likelihood})/n + \log(nh^*)/(nh^*) \times df_3,$$

where df_3 is the number of covariates with nonzero effect.

2.3 Theoretical properties

In this section, we establish the asymptotic properties of the initial estimators, and our proposed KGNG and KGNG2 estimators.

2.3.1 Asymptotic Properties of Initial Estimators

Denote the true mean and deviation part of $\boldsymbol{\beta}_0(t)$ as $\boldsymbol{m}_0 = (m_{01}, \ldots, m_{0p})^\intercal$ and $\boldsymbol{\beta}_0^*(t) = (\beta_{01}^*(t), \ldots, \beta_{0p}^*(t))^\intercal$, respectively, where

$$m_{0k} = \tau^{-1}\int_0^\tau \beta_{0k}(u)du, \quad \beta_{0k}^*(t) = \beta_{0k}(t) - m_{0k},$$

for $k = 1, \ldots, p$. Let $I(\boldsymbol{\beta}, t) = -\partial^2 L_{1n}(\boldsymbol{\beta}, t)/\partial \boldsymbol{\beta}^2 = n^{-1}\sum_{i=1}^n \int_0^\tau V(\boldsymbol{\beta}, s)K_{h_n}(s-t)dN_i(s)$, where

$$\begin{aligned} V(\boldsymbol{\beta}, t) =& \frac{S^{(2)}(\boldsymbol{\beta}, t)}{S^{(0)}(\boldsymbol{\beta}, t)} - \left(\frac{S^{(1)}(\boldsymbol{\beta}, t)}{S^{(0)}(\boldsymbol{\beta}, t)}\right)^{\otimes 2}, \\ S^{(r)}(\boldsymbol{\beta}, t) =& n^{-1}\sum_{i=1}^n Y_i(t)\boldsymbol{Z}_i^{\otimes r} e^{\boldsymbol{\beta}'\boldsymbol{Z}_i}, r = 0, 1, 2, \end{aligned}$$

with $\otimes$ denoting the outer product. Let $\boldsymbol{E}(\boldsymbol{\beta}, t) = S^{(1)}(\boldsymbol{\beta}, t)/S^{(0)}(\boldsymbol{\beta}, t)$, $P(t|\boldsymbol{z}) = \mathrm{P}(\widetilde{T} \geq t|\boldsymbol{Z} = \boldsymbol{z})$, $Q_0(t) = \mathrm{E}[P(t|\boldsymbol{Z})\alpha(t|\boldsymbol{Z})]$, $Q_1(t) = \mathrm{E}[P(t|\boldsymbol{Z})\alpha(t|\boldsymbol{Z})\boldsymbol{Z}]$ and $Q_2(t) = \mathrm{E}[P(t|\boldsymbol{Z})\alpha(t|\boldsymbol{Z})\boldsymbol{Z}^{\otimes 2}]$. Define $\Sigma(t) = Q_2(t) - Q_1(t)Q_1(t)^\intercal/Q_0(t)$. Let $s^{(r)}(\boldsymbol{\beta}, t)$ denote the limits of $S^{(r)}(\boldsymbol{\beta}, t)$, $r = 0, 1, 2$, as $n \to \infty$. Let $\mathcal{N}(t, \epsilon)$ be an ϵ-neighborhood of t, for $\epsilon > 0$ and $t \in [0, \tau]$, and $\mathcal{B}$ be a compact set of IR^p that includes a neighborhood of $\boldsymbol{\beta}_0(t)$ for $t \in [0, \tau]$. Under the following regularity conditions,

(A.1) The kernel function $K(\cdot)$ is a bounded and symmetric density with a bounded support [-1,1];

(A.2) For $t \in [0, \tau]$

$$\mathrm{E}\left[\exp\left\{2\left(\sup_{u\in\mathcal{N}(t,\epsilon)} |\boldsymbol{\beta}_0(u)| + \boldsymbol{\beta}_0'(t) + 3\right)|Z|\right\}\right] < \infty;$$

(A.3) $Q_0(u) > 0$, $Q_1(u)$ and $Q_2(u)$ are continuous for $t \in [0, \tau]$;

(A.4) Assume that $\alpha_0(t)$ is positive and continuous, $P(t|\boldsymbol{z}) > 0$ and coefficient functions $\{\beta_{0j}(t)\}$ have a continuous second derivative for $t \in [0, \tau]$;

(A.5) Assume that the matrix $\Sigma(t)$ is positive definite for $t \in [0, \tau]$;

(A.6) $s^{(r)}(\boldsymbol{\beta}, t)$ is uniformly continuous with respect to $(\boldsymbol{\beta}^\intercal, t)^\intercal \in \mathcal{B} \times [0, \tau]$ for $t \in [0, \tau]$;

we have

Lemma 2.1. *Suppose* $h_n = O(n^{-\nu})$ *with* $\nu \in [1/5, 1)$.

(a) $\widetilde{\boldsymbol{\beta}}(t) \xrightarrow{p} \boldsymbol{\beta}_0(t)$, $0 \le t \le \tau$.

(b) If $1/5 < \nu < 1$, *we have, for fixed* $t \in (0, \tau)$,

$$(nh_n)^{1/2}\left(\widetilde{\boldsymbol{\beta}}(t) - \boldsymbol{\beta}_0(t)\right) \xrightarrow{d} N\left\{0,\ \Sigma^{-1}(t)\int_{-1}^{1} K^2(s)ds\right\}.$$

where $\Sigma(t)$ *can be consistently estimated by* $I(\widetilde{\boldsymbol{\beta}}(t), t)$.

Lemma 2.2. *Suppose* $h_n = O(n^{-\nu})$ *with* $1/4 < \nu < 1/2$, *we have*

$$n^{1/2}\left(\widetilde{\boldsymbol{m}} - \boldsymbol{m}_0\right) \xrightarrow{d} N\left(0,\ \Sigma_m\right),$$

where $\Sigma_m = \int_0^\tau \Sigma^{-1}(u)du/\tau^2$ *and can be consistently estimated by*

$$\widehat{\Sigma}_m = \int_h^{\tau-h} I^{-1}(\widetilde{\beta}(u), u)du/\left(\tau - 2h\right)^2.$$

The proof of Lemma 2.1 follows Cai and Sun (2003) which is omitted. The proof of Lemma 2.2 follows similar steps given in Section 5 of Tian et al. (2005) and is relegated to the Supplement.

2.3.2 Asymptotic Properties of KGNG Estimators

Let $\boldsymbol{\lambda}_{01}$ and $\boldsymbol{\lambda}_{02}$ denote the true values of $\boldsymbol{\lambda}_1$ and $\boldsymbol{\lambda}_2$, respectively, and partition them into three parts: $\boldsymbol{\lambda}_{01} = (\boldsymbol{\lambda}_{01}^{O\,\intercal}, \boldsymbol{\lambda}_{01}^{C\,\intercal}, \boldsymbol{\lambda}_{01}^{NC\intercal})^\intercal$ and $(\boldsymbol{\lambda}_{02}^{O\,\intercal}, \boldsymbol{\lambda}_{02}^{C\,\intercal}, \boldsymbol{\lambda}_{02}^{NC\intercal})^\intercal$, according to the true index sets I_O, I_C and I_{NC}, respectively. Then

$$\boldsymbol{\lambda}_{01}^{O} = \mathbf{0}_{p_1}, \quad \boldsymbol{\lambda}_{01}^{C} = \mathbf{1}_{p_2}, \quad \boldsymbol{\lambda}_{02}^{O} = \mathbf{0}_{p_1}, \quad \boldsymbol{\lambda}_{02}^{C} = \mathbf{0}_{p_2}, \quad \boldsymbol{\lambda}_{02}^{NC} = \mathbf{1}_{p_3}$$

and $\lambda_{01,j}^{NC} = 0$ or 1 for $j \in \mathrm{I}_{NC}$, where $\lambda_{01,j}^{NC}$ is the jth component of $\boldsymbol{\lambda}_{01}^{NC}$ and corresponds to the mean part of the jth covariate with time-varying effect. If $\lambda_{01,j}^{NC} = 0$, the jth time-varying effect has zero mean effect; otherwise, it has nonzero mean effect. As we do not distinguish between the above two types of time-varying effects, without loss of generality, we assume $\boldsymbol{\lambda}_{01}^{NC} = \mathbf{1}_{p_3}$ in our theoretical derivations. Let $\boldsymbol{\lambda}_0 = (\boldsymbol{\lambda}_{01}^\intercal, \boldsymbol{\lambda}_{02}^\intercal)^\intercal$. We further partition $\boldsymbol{\lambda}_0$ as $\boldsymbol{\lambda}_0^{(1)}$ representing all the ones and $\boldsymbol{\lambda}_0^{(0)}$ representing all the zeros, where $\boldsymbol{\lambda}_0^{(1)} = (\boldsymbol{\lambda}_{01}^{C\,\intercal}, \boldsymbol{\lambda}_{01}^{NC\intercal}, \boldsymbol{\lambda}_{02}^{NC\intercal})^\intercal = \mathbf{1}_{p_2+2p_3}$ and $\boldsymbol{\lambda}_0^{(0)} = (\boldsymbol{\lambda}_{01}^{O\,\intercal}, \boldsymbol{\lambda}_{02}^{O\,\intercal}, \boldsymbol{\lambda}_{02}^{C\,\intercal})^\intercal = \mathbf{0}_{2p_1+p_2}$. In a similar manner we define $\boldsymbol{\lambda}$, $\boldsymbol{\lambda}^{(1)}$, $\boldsymbol{\lambda}^{(0)}$, $\widehat{\boldsymbol{\lambda}}$, $\widehat{\boldsymbol{\lambda}}^{(1)}$ and $\widehat{\boldsymbol{\lambda}}^{(0)}$, which accord to the true index sets I_O, I_C and I_{NC}.

Note that the KGNG estimator defined in (2.4) takes the form $\widehat{\boldsymbol{\beta}}(t) = \widehat{\boldsymbol{\lambda}}_1 \circ \widetilde{\boldsymbol{m}} + \widehat{\boldsymbol{\lambda}}_2 \circ \widetilde{\boldsymbol{\beta}}^*(t)$. To derive its asymptotic properties, we need to study the asymptotic properties of $\widetilde{\boldsymbol{m}}$, $\widetilde{\boldsymbol{\beta}}^*(t)$ and $\widehat{\boldsymbol{\lambda}}$. In Lemmas 2.1 and 2.2, we have established the asymptotic properties of $\widetilde{\boldsymbol{m}}$ and $\widetilde{\boldsymbol{\beta}}^*(t)$. In the following, we derive the asymptotic properties of $\widehat{\boldsymbol{\lambda}}$, which consists of $\widehat{\boldsymbol{\lambda}}^{(1)}$ and $\widehat{\boldsymbol{\lambda}}^{(0)}$. In particular, we establish the root-n consistency of $\widehat{\boldsymbol{\lambda}}^{(1)}$ in Theorem 2.1 and the sparsity property of $\widehat{\boldsymbol{\lambda}}^{(0)}$ in Theorem 2.2.

Theorem 2.1. *Assume $h_n = O(n^{-\nu})$ with $1/4 < \nu < 1/2$. Under the regularity conditions assumed in Lemmas 2.1 and 2.2, if $\max(\theta_1, \theta_2)/\sqrt{n}$ is bounded, then $\|\widehat{\boldsymbol{\lambda}}^{(1)} - \boldsymbol{\lambda}_0^{(1)}\| = O_p(n^{-1/2})$.*

Theorem 2.2. *Assume $h_n = O(n^{-\nu})$ with $1/4 < \nu < 1/2$. Under the regularity conditions assumed in Lemmas 2.1 and 2.2, if $\|\widehat{\boldsymbol{\lambda}}^{(1)} - \boldsymbol{\lambda}_0^{(1)}\| = O_p(n^{-1/2})$ and $h_n^{1/2} \min(\theta_1, \theta_2) \to \infty$, then $P(\widehat{\boldsymbol{\lambda}}^{(0)} = \mathbf{0}) \to 1$.*

Combining Theorems 2.1 and 2.2, we can then prove the selection consistency of the KGNG estimator, which is summarized in part (a) of Theorem 2.3. We further establish the asymptotic normality of the KGNG estimators for both nonzero constant and time-varying regression coefficients in Theorem 2.3. Recall that $\boldsymbol{\beta}_0(t) = (\boldsymbol{\beta}_O^\intercal, \boldsymbol{\beta}_C^\intercal, \boldsymbol{\beta}_{NC}^\intercal(t))^\intercal \in \mathbb{R}^p$, where $\boldsymbol{\beta}_O \in \mathbb{R}^{p_1}$, $\boldsymbol{\beta}_C \in \mathbb{R}^{p_2}$ and $\boldsymbol{\beta}_{NC}(t) \in \mathbb{R}^{p_3}$ are subvector of $\boldsymbol{\beta}_0(t)$ corresponding to the true underline index sets I_O, I_C and I_{NC}, respectively. In a similar manner, we write $\boldsymbol{m}_0 = (\boldsymbol{m}_O^\intercal, \boldsymbol{m}_C^\intercal, \boldsymbol{m}_{NC}^\intercal)^\intercal$ and $\boldsymbol{\beta}_0^*(t) = ({\boldsymbol{\beta}_O^*}^\intercal, {\boldsymbol{\beta}_C^*}^\intercal, {\boldsymbol{\beta}_{NC}^*}^\intercal(t))^\intercal$. We further partition $\widehat{\boldsymbol{\beta}}(t)$, $\widetilde{\boldsymbol{\beta}}(t)$, $\widetilde{\boldsymbol{m}}$ and $\widetilde{\boldsymbol{\beta}}^*(t)$ accordingly.

Theorem 2.3. *Assume $h_n = O(n^{-\nu})$ with $1/4 < \nu < 1/2$. Under the regularity conditions assumed in Lemmas 1 and 2, if $\max(\theta_1, \theta_2)/\sqrt{n}$ is bounded and $h_n^{1/2} \min(\theta_1, \theta_2) \to \infty$, then*

(a) ***(Selection consistency)*** *with probability tending to one, $\widehat{\mathrm{I}}_O = \mathrm{I}_O$, $\widehat{\mathrm{I}}_C = \mathrm{I}_C$ and $\widehat{\mathrm{I}}_{NC} = \mathrm{I}_{NC}$.*

(b) ***(Root-n consistency of $\widehat{\boldsymbol{\beta}}_C$)*** *$\widehat{\boldsymbol{\beta}}_C$ is a root-n consistent estimator for $\boldsymbol{\beta}_C$.*

(c) ***(Asymptotic normality of $\widehat{\boldsymbol{\beta}}_C$)*** *if we further assume $\max(\theta_1, \theta_2)/\sqrt{n} \to 0$,*

$$(n)^{1/2}\left(\widehat{\boldsymbol{\beta}}_C - \boldsymbol{\beta}_C\right) \xrightarrow{d} N\left\{0,\ \Sigma_m^F\right\},$$

where

$$\Sigma_m^F = \int_0^\tau \left(D(u) + \frac{1}{\tau} B_{10}\Sigma^{-1}(u)\right)\Sigma(u)\left(D(u) + \frac{1}{\tau} B_{10}\Sigma^{-1}(u)\right)^\intercal du. \tag{2.6}$$

Here $D(u)$ is a $p_2 \times p$ matrix given in (2.25) in the Supplement and $B_{10} = (\mathbf{0}_{p_2\times p_1} | \boldsymbol{I}_{p_2} | \mathbf{0}_{p_2\times p_3})$.

(d) ***(Asymptotic normality of*** $\widehat{\boldsymbol{\beta}}_{NC}(t)$***)***

$$(nh_n)^{1/2}\left(\widehat{\boldsymbol{\beta}}_{NC}(t)-\boldsymbol{\beta}_{NC}(t)\right)\xrightarrow{d}N\left\{0,\ \{\Sigma^{-1}(t)\}_{NC,NC}\int_{-1}^{1}K^2(s)ds\right\},$$

where $\{\Sigma^{-1}(t)\}_{NC,NC}$ *is the submatrix of* $\Sigma^{-1}(t)$ *corresponding to* I_{NC}.

The proofs of Theorem 2.1, 2.2 and 2.3 are relegated to the Supplement. The asymptotic variance-covariance matrices given in parts (c) and (d) can be consistently estimated by the usual plug-in method. It is interesting to note that the limiting distribution of the KGNG estimator for the time-varying coefficients given in part (d) is the same as that of the corresponding initial estimator. Actually in the proof of part (d), we have shown that the difference between $\widehat{\boldsymbol{\beta}}_{NC}(t)$ and $\widetilde{\boldsymbol{\beta}}_{NC}(t)$ are uniformly asymptotically negligible in t. Thus we could construct confidence bands of $\widehat{\boldsymbol{\beta}}_{NC}(t)$ using the resampling technique as proposed in Tian et al. (2005).

2.3.3 Asymptotic Properties of KGNG2 Estimator

Let $\boldsymbol{Z}=(\boldsymbol{Z}_O^\intercal,\boldsymbol{Z}_C^\intercal,\boldsymbol{Z}_{NC}^\intercal)^\intercal$ and $\underline{\boldsymbol{Z}}=(\boldsymbol{Z}_C^\intercal,\boldsymbol{Z}_{NC}^\intercal)^\intercal$ as the subvector of $\boldsymbol{Z}$ with only important covariates kept. Let $\underline{\boldsymbol{\beta}}_0(t)=(\boldsymbol{\beta}_C^\intercal(t),\boldsymbol{\beta}_{NC}^\intercal(t))^\intercal$, $\alpha(t|\underline{\boldsymbol{Z}})=\alpha_0(t)e^{\underline{\boldsymbol{\beta}}_0(t)^\intercal\underline{\boldsymbol{Z}}}$, $P(t|\underline{\boldsymbol{Z}})=P(Y\ge t|\underline{\boldsymbol{Z}}=\underline{\boldsymbol{z}})$, $\underline{Q}_0(t)=E[P(t|\underline{\boldsymbol{Z}})\alpha(t|\underline{\boldsymbol{Z}})]$, $\underline{Q}_1(t)=E[P(t|\underline{\boldsymbol{Z}})\alpha(t|\underline{\boldsymbol{Z}})\underline{\boldsymbol{Z}}]$ and $\underline{Q}_2(t)=E[P(t|\underline{\boldsymbol{Z}})\alpha(t|\underline{\boldsymbol{Z}})\underline{\boldsymbol{Z}}^{\otimes 2}]$. Define $\underline{\Sigma}(t)=\underline{Q}_2(t)-\underline{Q}_1(t)\underline{Q}_1(t)^\intercal/\underline{Q}_0(t)$. We add "*" to distinguish the KGNG2 estimator from the KGNG estimator. We summarize the asymptotic properties of KGNG2 estimator in Theorem 2.4.

Theorem 2.4. *Assume* $h_n=O(n^{-\nu})$ *with* $1/4<\nu<1/2$. *Under the regularity conditions assumed in Lemmas 1 and 2, if (1)* $\theta^*/\sqrt{n}$ *is bounded and* $h_n^{1/2}\theta^*\to\infty$*; (2)* $\max(\theta_1,\theta_2)/\sqrt{n}$ *is bounded and* $h_n^{1/2}\min(\theta_1,\theta_2)\to\infty$, *then*

(a) ***(Selection consistency of preliminary Step 0)*** *with probability tending to one,* $\widehat{\lambda}_k^*=0$, *for* $k\in\mathrm{I}_O$ *and* $\widehat{\lambda}_k^*\neq 0$, *for* $k\in\mathrm{I}_C\bigcup\mathrm{I}_{NC}$.

(b) ***(Selection consistency)*** *with probability tending to one,* $\widehat{\mathrm{I}}_O^*=\mathrm{I}_O$, $\widehat{\mathrm{I}}_C^*=\mathrm{I}_C$ *and* $\widehat{\mathrm{I}}_{NC}^*=\mathrm{I}_{NC}$.

(c) ***(Root-n consistency of*** $\widehat{\boldsymbol{\beta}}_C^*$***)*** $\widehat{\boldsymbol{\beta}}_C^*(t)$ *is a root-n consistent estimator for* $\boldsymbol{\beta}_C$*.*

(d) ***(Asymptotic normality of*** $\widehat{\boldsymbol{\beta}}_C^*$***)*** *if we further assume* $\max(\theta_1, \theta_2)/\sqrt{n} \to 0$,

$$(n)^{1/2}\left(\widehat{\boldsymbol{\beta}}_C^* - \boldsymbol{\beta}_C\right) \xrightarrow{d} N\left\{0,\ \underline{\Sigma}_m^F\right\},$$

where $\underline{\Sigma}_m^F$ *can be computed following* (2.6) *with some obvious changes.*

(e) ***(Asymptotic normality of*** $\widehat{\boldsymbol{\beta}}_{NC}^*(t)$***)***

$$(nh_n)^{1/2}\left(\widehat{\boldsymbol{\beta}}_{NC}^*(t) - \boldsymbol{\beta}_{NC}(t)\right) \xrightarrow{d} N\left\{0,\ \{\underline{\Sigma}^{-1}(t)\}_{NC,NC}\int_{-1}^{1}K^2(s)ds\right\},$$

where $\{\underline{\Sigma}^{-1}(t)\}_{NC,NC}$ *is the submatrix of* $\underline{\Sigma}^{-1}(t)$ *corresponding to* I_{NC}*.*

The proof of Theorem 2.4 is similar to that of Theorem 2.3 and hence is omitted here. Based on Theorem 2.3 and 2.4, $\widehat{\boldsymbol{\beta}}_{NC}^*(t)$ is strictly more efficient than $\widehat{\boldsymbol{\beta}}_{NC}(t)$ if I_O is not empty. However, there is no clear order between the efficiencies of $\widehat{\boldsymbol{\beta}}_C^*$ and $\widehat{\boldsymbol{\beta}}_C$.

2.4 Numerical studies

2.4.1 Simulation studies

We generate failure times from the varying-coefficient Cox model (2.1). Here, covariate vector $\boldsymbol{Z}$ is generated from a multivariate normal distribution with mean 0, variance 0.5 and correlation coefficient $0.5^{|j-k|}$ for any pair (j, k). We consider both the low dimensional case and the high dimensional case, where the dimension $p = 10$ and 50 respectively. There are three nonzero coefficients in $\boldsymbol{\beta}_0(t)$, i.e., $\beta_{02}(t) = -\{1 + \cos(\pi t)\}\mathbb{1}(0 < t < 1)$, $\beta_{03}(t) = 1.5\{\cos(\pi t/2)\}$ and $\beta_{08}(t) = -1$. So, there are two covariates with time-varying effects, one with constant effect and all the remains with null effect. The baseline hazard function $\alpha_0(t) = \exp\{-\cos(\pi t/2)\}$. We consider both the cases when censoring times are dependent and independent of the covariates. They are given in the cases $p = 10$ and $p = 50$, respectively. When $p = 10$, the censoring times of

ith subject is a mixture of W and a point mass at 2, where $W = \min(\exp(Z_{i2} - Z_{i5}),\ \mathrm{Unif}(0, 2))$. When $p = 50$, we generate censoring times from a mixture of $\mathrm{Unif}(0, 2))$ and a point mass at 2. For both cases, the mixing probability is chosen to have the desired censoring proportion, i.e. $c_p = 20\%$ or 40%. For each scenario, we conduct 100 simulation runs with sample size $n = 200$ and 400. We compare the proposed KGNG and KGNG2 with the AGLASSO of Yan and Huang (2012). For our estimators, we use the Epanechnikov kernel $K(x) = 3(1 - x^2)/4$, $-1 \le x \le 1$. The bandwidth h is chosen using 5-fold cross validation as discussed in Section 2.2.3. For KGNG2, the same bandwidth is used for Step 1 as for Step 0. We use the proposed BIC criterion in Section 2.2.3 to tune (θ_1, θ_2) and θ^*.

Tables 2.1 and 2.2 summarize the mean squared errors and variable selection results for $p = 10$ and 50, respectively. The selection frequency of each variable over 100 runs is reported, where the important covariates are Z_2, Z_3 and Z_8. The mean squared error (MSE) is calculated as

$$\frac{1}{100} \sum_{i=1}^{100} \left\{ \widehat{\boldsymbol{\beta}}(t_i) - \boldsymbol{\beta}_0(t_i) \right\}^{\mathsf{T}} V \left\{ \widehat{\boldsymbol{\beta}}(t_i) - \boldsymbol{\beta}_0(t_i) \right\},$$

where $\{t_1, \ldots, t_{100}\}$ are 100 equally-spaced grid points in the time interval $(0, 2)$ and V is the population covariance matrix of covariates.

From Tables 2.1 and 2.2, we make the following observations. First, KGNG2 shows the best performance in terms of variable selection and MSE in almost all scenarios, especially for large dimension case ($p = 50$). This is expected since KGNG2 is a two-stage approach, by first excluding the noise variables. Second, both KGNG and KGNG2 outperform AGLASSO with deduction in MSE as large as 60%. Third, KGNG and KGNG2 select Z_2, Z_3 and Z_8 as important covariates nearly all the times, while the AGLASSO misses Z_3 occasionally, especially when the sample size is small.

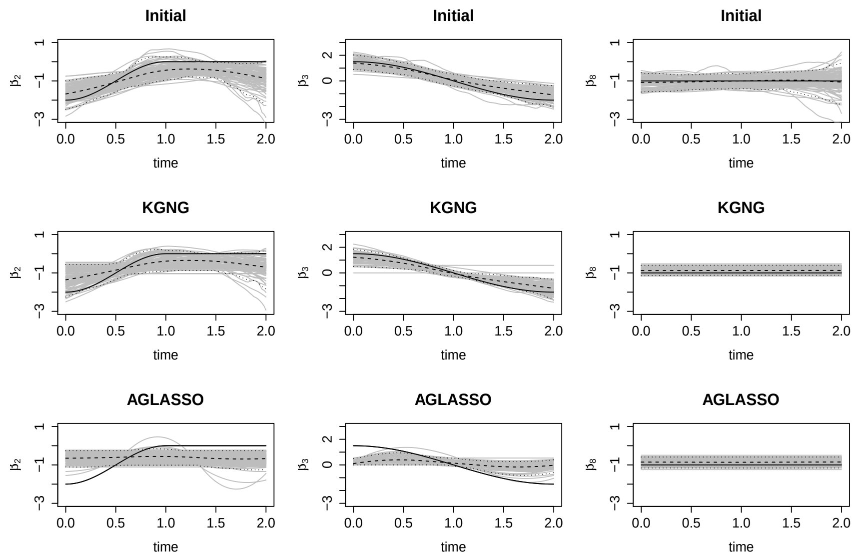
Table 2.3 summarizes the structure selection result of the three covariates with nonzero effect. We report the frequencies of each covariate being classified into the three categories I_O, I_C and I_{NC}. In summary, the proposed KGNG and KGNG2 outperform AGLASSO for all three covariates, exhibiting the most significant improvement for Z_2. The AGLASSO tends to falsely

Table 2.1: Variable selection and estimation results for $p = 10$. MSE stands for mean-squared error. Standard deviations of the Monte Carlo estimates are given in parentheses.

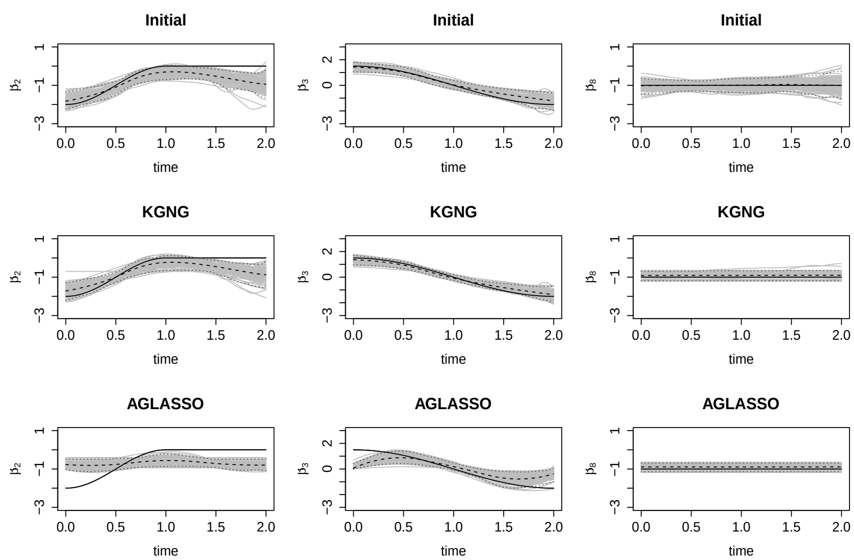
n	c_p	method	Z_1	Z_2	Z_3	Z_4	Z_5	Z_6	Z_7	Z_8	Z_9	Z_{10}	MSE (SD)
200	20	KGNG	7	100	99	9	5	10	10	100	11	7	0.180 (0.107)
		KGNG2	3	100	100	4	2	5	4	100	6	1	0.172 (0.126)
		AGLASSO	6	100	73	5	2	2	8	100	5	4	0.444 (0.103)
400	20	KGNG	1	100	100	4	2	3	6	100	4	3	0.114 (0.053)
		KGNG2	0	100	100	0	1	0	1	100	4	0	0.114 (0.054)
		AGLASSO	0	100	100	2	2	0	1	100	2	0	0.268 (0.055)
200	40	KGNG	10	100	98	14	10	9	11	100	14	5	0.244 (0.151)
		KGNG2	7	100	100	7	5	3	5	100	11	3	0.227 (0.176)
		AGLASSO	7	100	82	9	4	4	6	100	6	1	0.492 (0.122)
400	40	KGNG	5	100	100	3	3	4	4	100	7	2	0.117 (0.080)
		KGNG2	2	100	100	1	1	0	1	100	3	1	0.110 (0.126)
		AGLASSO	1	100	100	2	0	0	3	100	5	1	0.303 (0.096)

select the time-varying-effect Z_2 as a constant-effect covariate. For example, when $p = 10$, $n = 200$ and $c_p = 20\%$, AGLASSO correctly classifies Z_2 only 34 times out of 100, while both KGNG and KGNG2 classify Z_2 correctly for more than 70 times.

Finally, we plot the initial estimator, KGNG and AGLASSO for three nonzero coefficients and their pointwise 95% confidence intervals based on 100 simulations for $p = 10$. The plots for $c_p = 20\%$ and $c_p = 40\%$ are given in Figures 2.1 and 2.2, respectively. We make the following observations. First, the performance of both KGNG and AGLASSO improves substantially when sample size increases. Second, KGNG has smaller biases in the estimation of time-varying coefficients compared with AGLASSO in all the cases, while KGNG and AGLASSO give comparable estimates for the constant coefficient. Third, the improvement of KGNG over the initial estimator for time-varying coefficients is not obvious, however, the improvement for the constant coefficient is significant. This matches with our Theorem 2.3 parts (c) and (d).

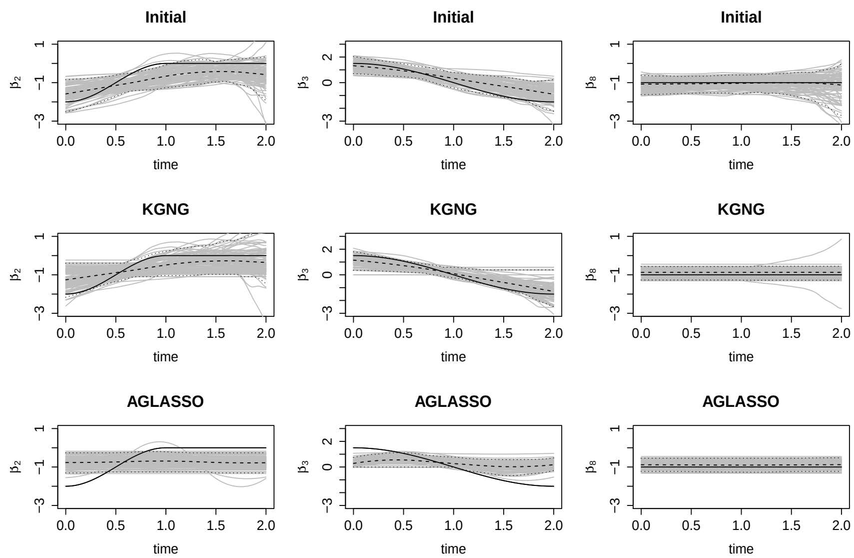


(a) n=200. Upper: Initial estimate. Middle: KGNG. Lower: AGLASSO.

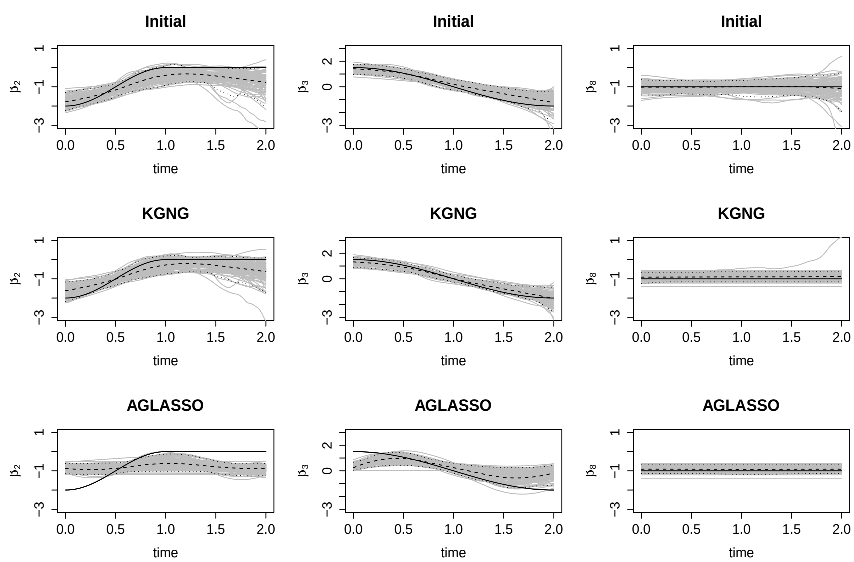


(b) n=400. Upper: Initial estimate. Middle: KGNG. Lower: AGLASSO.

Figure 2.1: Estimated curves (gray) of the three nonzero coefficients from 100 replicates when $c_p = 20\%$ and $p = 10$. The dark lines are the true curves. The dashed lines are the average of 100 estimates. The dotted lines are the simulation-based pointwise 95% confidence intervals.



(a) n=200. Upper: Initial estimate. Middle: KGNG. Lower: AGLASSO.



(b) n=400. Upper: Initial estimate. Middle: KGNG. Lower: AGLASSO.

Figure 2.2: Estimated curves (gray) of the three nonzero coefficients from 100 replicates when $c_p = 40\%$ and $p = 10$. The dark lines are the true curves. The dashed lines are the average of 100 estimates. The dotted lines are the simulation-based pointwise 95% confidence intervals.

Table 2.2: Variable selection and estimation results for $p = 50$. MSE stands for mean-squared error. Standard deviations of the Monte Carlo estimates are given in parentheses.

n	c_p	Method	Z_1	Z_2	Z_3	Z_4	Z_5	Z_6	Z_7	Z_8	Z_9	Z_{10}	$Z_{11} - Z_{50}$	MSE (SD)
200	20	KGNG	5	100	100	3	3	3	1	100	3	3	2.9	0.297 (0.095)
		KGNG2	3	99	87	3	3	3	3	100	3	2	2.4	0.291 (0.130)
		AGLASSO	10	100	79	4	5	3	7	100	7	2	3.3	0.425 (0.099)
400	20	KGNG	1	100	100	2	1	2	0	100	3	2	2.4	0.168 (0.056)
		KGNG2	0	100	100	2	0	1	0	100	1	0	0.6	0.134 (0.053)
		AGLASSO	3	100	100	2	1	1	4	100	0	2	1.2	0.261 (0.053)
200	40	KGNG	2	97	92	3	5	4	2	100	5	3	3.9	0.403 (0.147)
		KGNG2	2	98	87	4	3	2	2	100	5	1	2.0	0.367 (0.222)
		AGLASSO	12	100	91	12	11	6	10	100	11	9	6.2	0.494 (0.137)
400	40	KGNG	0	100	100	4	0	1	2	100	2	5	2.1	0.195 (0.058)
		KGNG2	0	100	100	1	0	0	1	100	1	0	1.3	0.181 (0.076)
		AGLASSO	5	100	100	4	1	2	3	100	6	2	4.3	0.292 (0.079)

2.4.2 Analysis of primary biliary cirrhosis (PBC) data

As an illustration, we apply our KGNG and KGNG2 methods to analyze the PBC data (Fleming and Harrington, 1991). The data is from the Mayo Clinic trial in primary biliary cirrhosis (PBC) of the liver conducted between 1974 and 1984. The primary biliary cirrhosis is a chronic disease in which the bile ducts in one's liver are slowly destroyed. In the study, 312 out of 424 patients who participate in the randomized trial are eligible for the analysis. There are 17 covariates: trtmt=treatment (Yes/No), age (in 10 years), gender=female/male, ascites=presence of ascites (Yes/No), hypato=presence of hepatomegaly (Yes/No), spiders=presence of spiders, edema=severity of oedema (0 denotes no oedema, 0.5 denotes untreated or successfully treated oedema and 1 denotes unsuccessfully treated oedema), logbili=logarithm of serum bilirubin (mg/dl), chol=serum cholesterol (mg/dl), logalb=logarithm of albumin (gm/dl), copper=urine copper (mg/day), alk=alkaline phosphatase (U/liter), sgot=liver enzyme (U/ml), trig=triglicerides (mg/dl), platelet=platelets per 10^{-3} ml^3, logprotime=logarithm of prothrombin time (seconds), stage=histologic stage of disease (category: 1, 2, 3 or 4).

The model selection of this data set has been previously studied in the context of both

Table 2.3: Structure selection results for covariates 2, 3 and 8. Here O is for null effect, C for constant effect, and NC for time-varying effect.

			p=10									**p=50**								
			Z_2			Z_3			Z_8			Z_2			Z_3			Z_8		
n	c_p	method	O	C	NC	O	C	NC	O	C	NC	O	C	NC	O	C	NC	O	C	NC
200	20	KGNG	0	21	79	1	1	98	0	97	3	0	62	38	0	1	99	0	99	1
		KGNG2	0	10	90	0	1	99	0	94	6	1	38	61	13	0	87	0	95	5
		AGLASSO	0	66	34	27	2	71	0	86	14	0	56	44	21	0	79	0	92	8
400	20	KGNG	0	1	99	0	0	100	0	97	3	0	1	99	0	0	100	0	97	3
		KGNG2	0	1	99	0	0	100	0	91	9	0	0	100	0	0	100	0	97	3
		AGLASSO	0	51	49	1	0	99	0	96	4	0	41	59	0	0	100	0	95	5
200	40	KGNG	0	31	69	2	4	94	0	97	3	3	57	40	8	11	81	0	98	2
		KGNG2	0	23	77	0	5	95	0	89	11	2	27	71	13	1	86	0	92	8
		AGLASSO	0	75	25	18	10	72	0	84	16	0	50	50	9	9	82	0	91	9
400	40	KGNG	0	0	100	0	0	100	0	97	3	0	16	84	0	0	100	0	99	1
		KGNG2	0	0	100	0	0	100	0	95	5	0	5	95	0	0	100	0	98	2
		AGLASSO	0	35	65	0	0	100	0	94	6	0	4	96	0	0	100	0	94	6

Cox model with time-independent coefficients (Tibshirani, 1997; Zhang and Lu, 2007) and Cox model with time-varying coefficients (Tian et al., 2005; Yan and Huang, 2012). To ease the comparison, we analyzed the data of 276 patients with no missingness in covariates and took log transformation of serum bilirubin, albumin and prothrombin time as in Yan and Huang (2012). For our methods, we used the 10-fold cross validation to find the optimal bandwidth in the initial estimator, which is 2000 (days). We chose $\tau = 3200$, which covers around 90% of observed survival times. Table 2.4 gives the estimates of coefficients by five methods: the maximum partial likelihood estimator (MPLE), the adaptive LASSO (ALASSO) estimator of Zhang and Lu (2007) based on a standard Cox model, the AGLASSO, and KGNG and KGNG2 based on a time-varying coefficient Cox model. The numbers given in parenthesis are the estimated standard errors for important constant coefficients selected by each method. The results for ALASSO and AGLASSO are copied directly from Yan and Huang (2012). We make the following observations. First, ALASSO, AGLASSO, KGNG and KGNG2 all select the same 7 important covariates: age, cooper, edema, logbili, logalb, logprotime and stage. Second, KGNG identifies three covariates with time-varying coefficients and KGNG2 identifies two, in which edema and logprotime are the common covariates identified by both KGNG and KGNG2. On the other hand, AGLASSO only selects logbilli as the covariate with time-varying coefficient. These results partly agree with the findings of Tian et al. (2005), where only 5 covariates: age, edema, logbili, logalb and logprotime, are considered in their time-varying coefficient Cox model, and three covariates: edema, logprotime and logbili, are identified as having time-varying coefficients. Finally, in Figures 2.3 and 2.4, we plotted the estimated coefficients by the initial step (maximum local partial likelihood estimator), KGNG and KGNG2 for the 7 important covariates and their associated 95% pointwise confidence intervals and simultaneous confidence bands.

Table 2.4: Analysis results for PBC data. TV stands for time-varying coefficients.

Covariate	MPLE	ALASSO	AGLASSO	KGNG	KGNG2
trtmt	-0.062 (0.211)				
age	0.261 (0.113)	0.270 (0.124)	0.263 (0.126)	0.211 (0.111)	0.253 (0.113)
gender	-0.256 (0.317)				
ascites	0.162 (0.381)				
hypato	-0.100 (0.254)				
spiders	0.049 (0.243)				
edema	0.926 (0.378)	0.842 (0.410)	0.932 (0.443)	TV	TV
logbili	0.723 (0.162)	0.699 (0.115)	TV	0.716 (0.096)	0.718 (0.116)
chol	0.000 (0.000)				
logalb	-2.270 (0.947)	-2.538 (0.762)	-2.440(0.789)	-2.173 (0.934)	-2.294 (1.183)
copper	1.694 (1.251)	2.218 (1.236)	2.089(1.261)	TV	1.379 (1.419)
alk	0.000 (0.000)				
sgot	0.003 (0.002)				
trig	-0.002 (0.001)				
platelet	0.001 (0.001)				
logprotime	2.335 (1.321)	2.099 (1.241)	1.822 (1.249)	TV	TV
stage	0.381 (0.176)	0.274 (0.140)	0.278 (0.143)	0.244 (0.110)	0.218 (0.130)

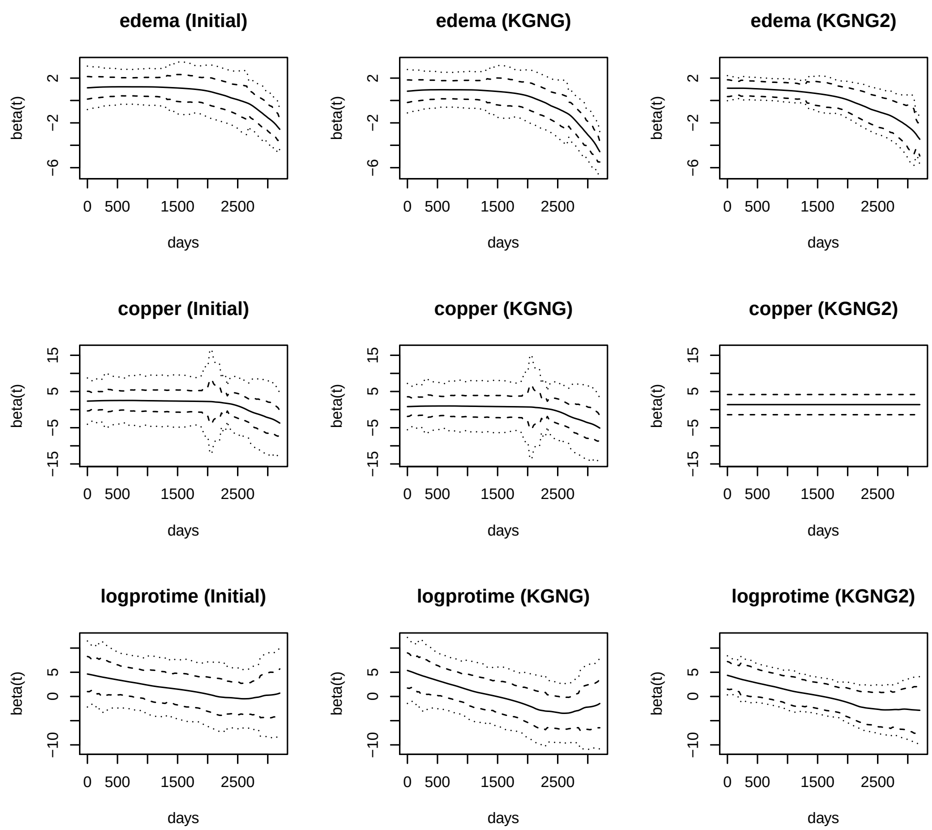


Figure 2.3: Estimated coefficients for covariates: edema, copper, and logprotime. Left panel: initial estimator in KGNG; Middle panel: KGNG; Right panel: KGNG2. Solid lines: estimated curves; Dashed lines: 95% pointwise confidence intervals; Dotted lines: 95% simultaneous confidence bands.

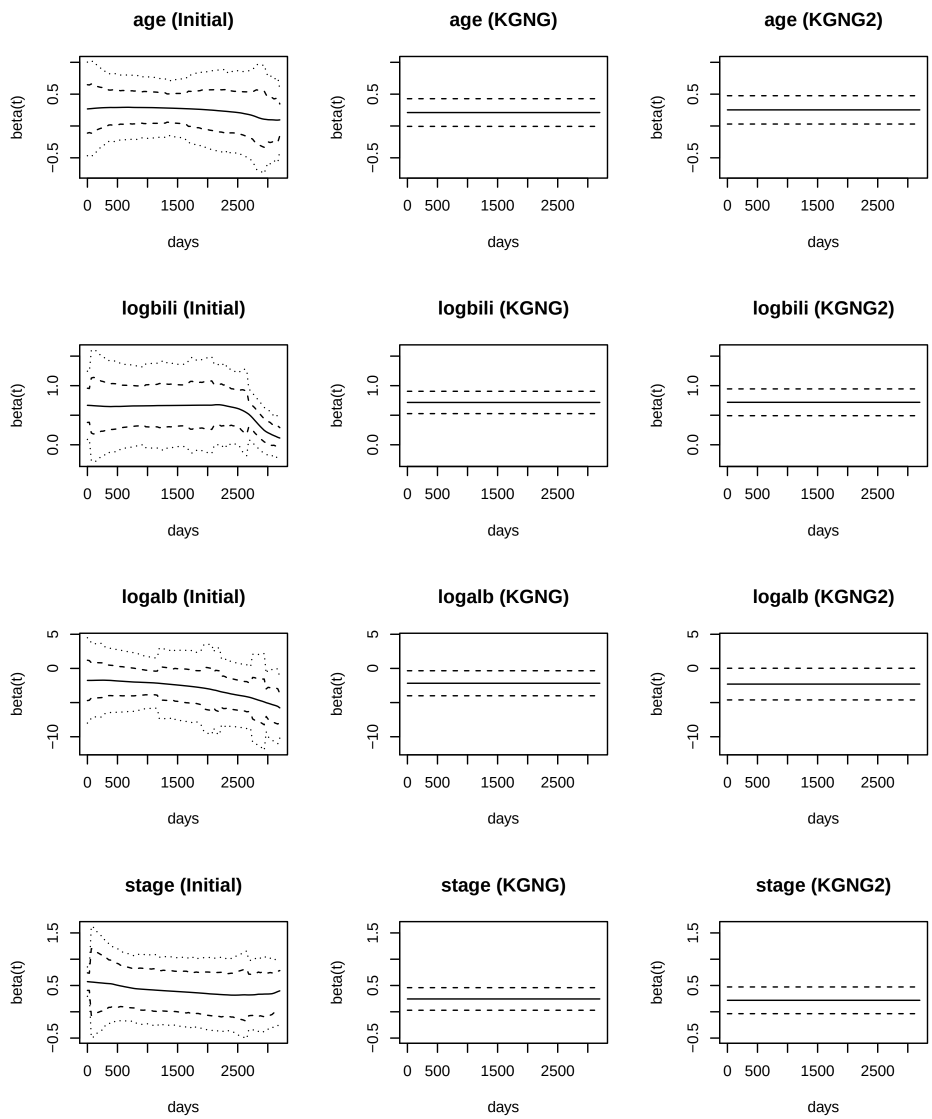


Figure 2.4: Estimated coefficients for covariates: age, logbili, logalb and stage. Left panel: initial estimator in KGNG; Middle panel: KGNG; Right panel: KGNG2. Solid lines: estimated curves; Dashed lines: 95% pointwise confidence intervals; Dotted lines: 95% simultaneous confidence bands.

2.5 Discussion

We propose a kernel group nonnegative garrote (KGNG) estimation method and its variant (KGNG2) for automatic structure selection and coefficient estimation in a time-varying coefficient Cox model. We establish the asymptotic properties, including structure selection consistency and asymptotic distributions, of our estimators for both constant and time-varying coefficients. Numerical studies have shown the competitive performance of the proposed methods compared with existing approaches.

In this chapter, we only focus on the case with fixed dimension p, and p is smaller than the sample size n. For $p > n$ case, a penalty term needs to be added to (2.2) to get reasonable initial estimates of the coefficient functions. Then Step 2 and 3 can follow similarly as proposed in this chapter. For an ultra-high dimension case, say, $p \gg n$, a screening procedure can be utilized to remove the noisy covariates beforehand. Then the dimension of the model can be decreased to a value that can be handled directly. However, a screening procedure for time-varying coefficient Cox model has not yet been developed in the literature, and it needs further investigation.

Since the proposed procedure depends on a large number of tuning parameters, it may be useful to develop a statistical test to check the goodness-of-fit of the final estimated model. We think that a cumulative sums of martingale residuals-based goodness-of-fit test can be derived for the final estimated model, following the techniques of Lin, Wei and Ying (1993). A similar goodness-of-fit test procedure was developed for the Dantzig selector in Cox's proportional hazards model (Antoniadis, Fryzlewicz and Letue, 2010). This is an interesting topic that needs further investigation.

Another interesting problem, as suggested by a referee, is to extend the proposed methods to domain selection, i.e., to develop a procedure which estimates the coefficients as exactly zero on parts of the time domain, and as nonzero (and time-varying) on the remaining parts. One simple solution would be to chop the coefficient functions evenly into small pieces on the study time domain and then apply the group nonnegative garrote penalty to identify the significant pieces. However, when there are a large number of covariates, this approach may be computationally challenging. The proposed KGNG/KGNG2 methods for structure selection can be regarded as a preliminary step for domain selection since they can help to remove all the covariates with null or constant effects and thus achieve effective dimension reduction. Then, the domain selection can only focus on the selected covariates with truly time-varying coefficients. This is an interesting extension that warrants our future research.

2.6 Supplement

This supplementary section contains the proofs of Lemma 2.2, and Theorem 2.1, 2.2 and 2.3.

Before we give the details of proofs, we first introduce some additional notation. With simple matrix multiplication, we have $\boldsymbol{\lambda}_0^{(1)} = P_1\boldsymbol{\lambda}_0$ and $\boldsymbol{\lambda}_0^{(0)} = P_2\boldsymbol{\lambda}_0$, where

$$P_1 = \begin{pmatrix} \mathbf{0}_{p_2\times p_1} & \boldsymbol{I}_{p_2} & \mathbf{0}_{p_2\times p_3} & \mathbf{0}_{p_2\times p_1} & \mathbf{0}_{p_2\times p_2} & \mathbf{0}_{p_2\times p_3} \\ \mathbf{0}_{p_3\times p_1} & \mathbf{0}_{p_3\times p_2} & \boldsymbol{I}_{p_3} & \mathbf{0}_{p_3\times p_1} & \mathbf{0}_{p_3\times p_2} & \mathbf{0}_{p_3\times p_3} \\ \mathbf{0}_{p_3\times p_1} & \mathbf{0}_{p_3\times p_2} & \mathbf{0}_{p_3\times p_3} & \mathbf{0}_{p_3\times p_1} & \mathbf{0}_{p_3\times p_2} & \boldsymbol{I}_{p_3} \end{pmatrix}_{(p_2+2p_3)\times 2p}; \tag{2.7}$$

$$P_2 = \begin{pmatrix} \boldsymbol{I}_{p_1} & \mathbf{0}_{p_1\times p_2} & \mathbf{0}_{p_1\times p_3} & \mathbf{0}_{p_1\times p_1} & \mathbf{0}_{p_1\times p_2} & \mathbf{0}_{p_1\times p_3} \\ \mathbf{0}_{p_1\times p_1} & \mathbf{0}_{p_1\times p_2} & \mathbf{0}_{p_1\times p_3} & \boldsymbol{I}_{p_1} & \mathbf{0}_{p_1\times p_2} & \mathbf{0}_{p_1\times p_3} \\ \mathbf{0}_{p_2\times p_1} & \mathbf{0}_{p_2\times p_2} & \mathbf{0}_{p_2\times p_3} & \mathbf{0}_{p_2\times p_1} & \boldsymbol{I}_{p_2} & \mathbf{0}_{p_2\times p_3} \end{pmatrix}_{(2p_1+p_2)\times 2p}. \tag{2.8}$$

Let $\overline{\boldsymbol{\lambda}}_0 = (\boldsymbol{\lambda}_0^{(1)^\top}, \boldsymbol{\lambda}_0^{(0)^\top})^\top$. We also have $\boldsymbol{\lambda}_{01} = P_3\overline{\boldsymbol{\lambda}}_0 = P_{31}\boldsymbol{\lambda}^{(1)} + P_{32}\boldsymbol{\lambda}^{(0)}$ and $\boldsymbol{\lambda}_{02} = P_4\overline{\boldsymbol{\lambda}}_0 =$

$P_{41}\boldsymbol{\lambda}^{(1)} + P_{42}\boldsymbol{\lambda}^{(0)}$, where

$$P_3 = (P_{31}|P_{32}) = \left(\begin{array}{ccc|ccc} \mathbf{0}_{p_1\times p_2} & \mathbf{0}_{p_1\times p_3} & \mathbf{0}_{p_1\times p_3} & \boldsymbol{I}_{p_1} & \mathbf{0}_{p_1\times p_1} & \mathbf{0}_{p_1\times p_2} \\ \boldsymbol{I}_{p_2} & \mathbf{0}_{p_2\times p_3} & \mathbf{0}_{p_2\times p_3} & \mathbf{0}_{p_2\times p_1} & \mathbf{0}_{p_2\times p_1} & \mathbf{0}_{p_2\times p_2} \\ \mathbf{0}_{p_3\times p_3} & \boldsymbol{I}_{p_3} & \mathbf{0}_{p_3\times p_3} & \mathbf{0}_{p_3\times p_1} & \mathbf{0}_{p_3\times p_1} & \mathbf{0}_{p_3\times p_2} \end{array}\right)_{p\times 2p}; \tag{2.9}$$

$$P_4 = (P_{41}|P_{42}) = \left(\begin{array}{ccc|ccc} \mathbf{0}_{p_1\times p_2} & \mathbf{0}_{p_1\times p_3} & \mathbf{0}_{p_1\times p_3} & \mathbf{0}_{p_1\times p_1} & \boldsymbol{I}_{p_1} & \mathbf{0}_{p_1\times p_2} \\ \mathbf{0}_{p_2\times p_2} & \mathbf{0}_{p_2\times p_3} & \mathbf{0}_{p_2\times p_3} & \mathbf{0}_{p_2\times p_1} & \mathbf{0}_{p_2\times p_1} & \boldsymbol{I}_{p_2} \\ \mathbf{0}_{p_3\times p_2} & \mathbf{0}_{p_3\times p_3} & \boldsymbol{I}_{p_3} & \mathbf{0}_{p_3\times p_1} & \mathbf{0}_{p_3\times p_1} & \mathbf{0}_{p_3\times p_2} \end{array}\right)_{p\times 2p}. \tag{2.10}$$

Let $L_{2n}(\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2; \widetilde{\boldsymbol{m}}, \widetilde{\boldsymbol{\beta}}^*(\cdot))$ and $Q_{2n}(\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2; \widetilde{\boldsymbol{m}}, \widetilde{\boldsymbol{\beta}}^*(\cdot))$ denote the partial likelihood and penalized partial likelihood in equation (2.3) respectively. We then have

$$\begin{aligned} L_{2n}(\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2; \boldsymbol{m}, \boldsymbol{\beta}(\cdot)) =& \sum_{i=1}^n \int_0^\tau \left[a(s)^\intercal \boldsymbol{Z}_i - \log\left(\sum_{j=1}^n Y_j(s) e^{a(s)^\intercal \boldsymbol{Z}_j} \right) \right] dN_i(s); \\ Q_{2n}(\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2; \boldsymbol{m}, \boldsymbol{\beta}(\cdot)) =& - L_{2n}(\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2; \boldsymbol{m}, \boldsymbol{\beta}(\cdot)) + \theta_1 \|\boldsymbol{\lambda}_1\|_1 + \theta_2 \|\boldsymbol{\lambda}_2\|_1, \end{aligned}$$

where

$$a(s) \equiv a(s, \boldsymbol{\lambda}^{(1)}, \boldsymbol{\lambda}^{(0)}; \boldsymbol{m}, \boldsymbol{\beta}^*(\cdot)) \equiv a(s, \boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2; \boldsymbol{m}, \boldsymbol{\beta}^*(\cdot)) = \boldsymbol{\lambda}_1 \circ \boldsymbol{m} + \boldsymbol{\lambda}_2 \circ \boldsymbol{\beta}^*(s). \tag{2.11}$$

Let $U(\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2; \boldsymbol{m}, \boldsymbol{\beta}^*(\cdot))$ and $H(\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2; \boldsymbol{m}, \boldsymbol{\beta}^*(\cdot))$ denote the vector and matrix of the first and second order partial derivatives of $L_{2n}(\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2; \boldsymbol{m}, \boldsymbol{\beta}(\cdot))$ with respect to $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_1^\intercal, \boldsymbol{\lambda}_2^\intercal)^\intercal$. Thus

$$\begin{aligned} U(\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2; \boldsymbol{m}, \boldsymbol{\beta}(\cdot)) =& \sum_{i=1}^n \int_0^\tau A(s) \left(\boldsymbol{Z}_i - \boldsymbol{E}(a(s), s) \right) dN_i(s), \\ H(\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2; \boldsymbol{m}, \boldsymbol{\beta}(\cdot)) =& - \sum_{i=1}^n \int_0^\tau A(s) \left[\frac{S^{(2)}\left(a(s), s\right)}{S^{(0)}\left(a(s), s\right)} - \left(\frac{S^{(1)}\left(a(s), s\right)}{S^{(0)}\left(a(s), s\right)} \right)^{\otimes 2} \right] A(s)^\intercal dN_i(s), \end{aligned}$$

where

$$A(s) \equiv A(s; \boldsymbol{m}, \boldsymbol{\beta}^*(\cdot)) = \begin{pmatrix} \mathrm{diag}(\boldsymbol{m}) \\ \mathrm{diag}(\boldsymbol{\beta}^*(s)) \end{pmatrix}_{2p \times p}. \tag{2.12}$$

Rearranging the order of $\boldsymbol{\lambda}$, we rewrite $L_{2n}(\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2; \cdot)$ as $L_{2n}(\boldsymbol{\lambda}^{(1)}, \boldsymbol{\lambda}^{(0)}; \cdot)$. Denote the first and second order partial derivatives of $L_{2n}(\boldsymbol{\lambda}^{(1)}, \boldsymbol{\lambda}^{(0)}; \cdot)$ with respect to $\boldsymbol{\lambda}^{(1)}$ as $U_1(\boldsymbol{\lambda}^{(1)}, \boldsymbol{\lambda}^{(0)}; \cdot)$ and $H_{11}(\boldsymbol{\lambda}^{(1)}, \boldsymbol{\lambda}^{(0)}; \cdot)$. Denote the first and second order partial derivatives of $L_{2n}(\boldsymbol{\lambda}^{(1)}, \boldsymbol{\lambda}^{(0)}; \cdot)$ with respect to $\boldsymbol{\lambda}^{(0)}$ as $U_2(\boldsymbol{\lambda}^{(1)}, \boldsymbol{\lambda}^{(0)}; \cdot)$ and $H_{22}(\boldsymbol{\lambda}^{(1)}, \boldsymbol{\lambda}^{(0)}; \cdot)$. In the following proofs, we use $\|\cdot\|_1$ and $\|\cdot\|$ to denote ℓ_1 and ℓ_2 norms, respectively.

2.6.1 Proof of Lemma 2.2

We follow similar steps to prove Lemma 2.2 as in Section 5 of Tian et al. (2005). Denote $M_i(s) = N_i(s) - \int_0^s Y_i(u) a(u|\boldsymbol{Z}_i) du$ for all $i = 1, \ldots, p$. Thus $M_i(s)$ is a local square integrable martingale with respect to the filtration $\mathcal{F}_t = \{\boldsymbol{Z}_i, N_i(u), Y_i(u), i = 1, \ldots, n, 0 \le u \le t\}$. For $h_n = O(n^{-\nu})$ with $1/4 < \nu < 1/2$, it follows from Tian et al. (2005) that

$$\begin{aligned}
&\sqrt{n}(\widetilde{\boldsymbol{m}} - \boldsymbol{m}_0) = \frac{1}{\tau} \int_0^\tau \sqrt{n} \left(\widetilde{\boldsymbol{\beta}}(s) - \boldsymbol{\beta}_0(s) \right) ds \\
\approx& \frac{1}{\tau} \int_0^\tau (h_n)^{-1/2} \Sigma^{-1}(s) (n h_n)^{-1/2} \\
&\sum_{i=1}^n \int_0^\tau \left(\boldsymbol{Z}_i - \boldsymbol{E}\left(\boldsymbol{\beta}_0(s), u\right) \right) K\left(\frac{u-s}{h_n} \right) dM_i(u) ds \\
=& \frac{1}{\tau} n^{-1/2} \sum_{i=1}^n \int_0^\tau \left[\int_0^\tau \Sigma^{-1}(s) \left(\boldsymbol{Z}_i - \boldsymbol{E}\left(\boldsymbol{\beta}_0(s), u\right) \right) \frac{1}{h_n} K\left(\frac{u-s}{h_n} \right) ds \right] dM_i(u) \\
=& \frac{1}{\tau} n^{-1/2} \sum_{i=1}^n \int_0^\tau \Sigma^{-1}(u) \left(\boldsymbol{Z}_i - \boldsymbol{E}\left(\boldsymbol{\beta}_0(u), u\right) \right) dM_i(u) + o_p(1).
\end{aligned}$$

The first inequality sign in the above equation follows directly from (9.1) of Tian et al. (2005). Then by the martingale central limit theorem (Andersen and Gill, 1982), $n^{1/2}(\widetilde{\boldsymbol{m}} - \boldsymbol{m}_0)$ converges weakly to a normal distribution with mean 0 and variance V, where

$$
\begin{aligned}
V =&\frac{1}{n\tau^2}\sum_{i=1}^{n}\int_0^\tau \Sigma^{-1}(u)\left(\boldsymbol{Z}_i - \boldsymbol{E}\left(\boldsymbol{\beta}_0(u),u\right)\right)^{\otimes 2}\Sigma^{-1}(u)Y_i(u)\alpha(u|\boldsymbol{Z}_i)du \\
=&\frac{1}{\tau^2}\int_0^\tau \Sigma^{-1}(u)\left[\frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{Z}_i - \boldsymbol{E}\left(\boldsymbol{\beta}_0(u),u\right)\right)^{\otimes 2}Y_i(u)\alpha(u|\boldsymbol{Z}_i)\right]\Sigma^{-1}(u)du.
\end{aligned}
$$

By the central limit theorem for Banach space (Ledoux and Talagrand, 1991), the sandwich part

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{Z}_i - \boldsymbol{E}\left(\boldsymbol{\beta}_0(u),u\right)\right)^{\otimes 2}Y_i(u)\alpha(u|\boldsymbol{Z}_i) \\
=&\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{Z}_i^{\otimes 2}Y_i(u)\alpha(u|\boldsymbol{Z}_i) - \frac{2}{n}\sum_{i=1}^{n}\frac{S^{(1)}(\boldsymbol{\beta}_0(u),u)\boldsymbol{Z}_iY_i(u)\alpha(u|\boldsymbol{Z}_i)}{S^{(0)}(\boldsymbol{\beta}_0(u),u)} \\
&+\left(\boldsymbol{E}\left(\boldsymbol{\beta}_0(u),u\right)\right)^{\otimes 2}\frac{1}{n}\sum_{i=1}^{n}Y_i(u)\alpha(u|\boldsymbol{Z}_i) \\
=&Q_2(u) - \frac{2Q_1(u)Q_1^\intercal(u)}{Q_0(u)} + \frac{Q_1(u)Q_1^\intercal(u)}{\left(Q_0(u)\right)^2}Q_0(u) + O_p(n^{-1/2}) \\
=&\Sigma(u) + O_p(n^{-1/2}).
\end{aligned}
$$

From the above equation, we can show that $V \xrightarrow{p} \Sigma_m = \int_0^\tau \Sigma^{-1}(u)du/\tau^2$. Thus we finish the proof of part (a).

For part (b), we notice that the convergence rate of $\widetilde{\boldsymbol{m}}$ is $\sqrt{n}$ which is faster than that of $\widetilde{\boldsymbol{\beta}}(\cdot)$. Thus the conclusion of part (b) follows immediately.

2.6.2 Proof of Theorem 2.1

Define $\widetilde{a}(s) \equiv a(s, \boldsymbol{\lambda}^{(1)}, \boldsymbol{\lambda}^{(0)}; \widetilde{\boldsymbol{m}}, \widetilde{\boldsymbol{\beta}}^*(\cdot))$, $\widetilde{A}(s) \equiv A(s; \widetilde{\boldsymbol{m}}, \widetilde{\boldsymbol{\beta}}^*(\cdot))$ and

$$
A_0(s) \equiv A(s; \boldsymbol{m}_0, \boldsymbol{\beta}_0^*(\cdot)), \tag{2.13}
$$

where $a(s,\boldsymbol{\lambda}^{(1)},\boldsymbol{\lambda}^{(0)};\boldsymbol{m},\boldsymbol{\beta}^*(\cdot))$ and $A(s;\boldsymbol{m},\boldsymbol{\beta}^*(\cdot))$ are defined in (2.11) and (2.12), respectively. The second order partial derivative of $Q_{2n}(\boldsymbol{\lambda}^{(1)},\boldsymbol{\lambda}^{(0)};\widetilde{\boldsymbol{m}},\widetilde{\boldsymbol{\beta}}^*(\cdot))$ with respect to $\boldsymbol{\lambda}^{(1)}$ is

$$
\begin{aligned}
&\frac{\partial^2 Q_{2n}(\boldsymbol{\lambda}^{(1)},\boldsymbol{\lambda}^{(0)};\widetilde{\boldsymbol{m}},\widetilde{\boldsymbol{\beta}}^*(\cdot))}{\partial\boldsymbol{\lambda}^{(1)}\partial\boldsymbol{\lambda}^{(1)\top}} = -H_{11}(\boldsymbol{\lambda}^{(1)},\boldsymbol{\lambda}^{(0)};\widetilde{\boldsymbol{m}},\widetilde{\boldsymbol{\beta}}^*(\cdot))\\
=&\sum_{i=1}^{n}\int_0^\tau P_1\widetilde{A}(s)\left[\frac{S^{(2)}(\widetilde{a}(s),s)}{S^{(0)}(\widetilde{a}(s),s)}-\left(\frac{S^{(1)}(\widetilde{a}(s),s)}{S^{(0)}(\widetilde{a}(s),s)}\right)^{\otimes 2}\right]\left(P_1\widetilde{A}(s)\right)^\top dN_i(s)\\
=&\sum_{i=1}^{n}\int_0^\tau P_1\widetilde{A}(s)\left[\frac{\sum_{i<j}Y_i(s)Y_j(s)e^{\widetilde{a}(s)^\top\boldsymbol{Z}_i}e^{\widetilde{a}(s)^\top\boldsymbol{Z}_j}\,(\boldsymbol{Z}_i-\boldsymbol{Z}_j)^{\otimes 2}}{\left(S^{(0)}(\widetilde{a}(s),s)\right)^2}\right]\left(P_1\widetilde{A}(s)\right)^\top dN_i(s),
\end{aligned}
$$

which is positive definite for large n and P_1 is defined in (2.7). Thus, for any fixed $\boldsymbol{\lambda}^{(0)}$, there exists a unique minimizer $\widehat{\boldsymbol{\lambda}}^{(1)}$ of $Q_{2n}(\boldsymbol{\lambda}^{(1)},\boldsymbol{\lambda}^{(0)};\widetilde{\boldsymbol{m}},\widetilde{\boldsymbol{\beta}}^*(\cdot))$ when n is large.

Consider the C-ball $B_n(C)=\{\boldsymbol{\lambda}^{(1)}:\boldsymbol{\lambda}^{(1)}=\boldsymbol{\lambda}_0^{(1)}+n^{-1/2}\boldsymbol{u},\|\boldsymbol{u}\|\le C\}$, $C>0$, and denote its boundary by $\partial B_n(C)$. To prove $\|\widehat{\boldsymbol{\lambda}}^{(1)}-\boldsymbol{\lambda}_0^{(1)}\|=O_p(n^{-1/2})$, it is sufficient to show that, for any given $\epsilon>0$, there exists a large constant C so that

$$
\mathrm{pr}\left\{\sup_{\boldsymbol{\lambda}^{(1)}\in\partial B_n(C)} Q_{2n}(\boldsymbol{\lambda}^{(1)},\boldsymbol{\lambda}^{(0)};\widetilde{\boldsymbol{m}},\widetilde{\boldsymbol{\beta}}^*(\cdot)) < Q_{2n}(\boldsymbol{\lambda}_0^{(1)},\boldsymbol{\lambda}^{(0)};\widetilde{\boldsymbol{m}},\widetilde{\boldsymbol{\beta}}^*(\cdot))\right\}\le 1-\epsilon. \tag{2.14}
$$

Then we have

$$
\begin{aligned}
&D_{2n}(u)=Q_{2n}(\boldsymbol{\lambda}_0^{(1)}+n^{-1/2}\boldsymbol{u},\boldsymbol{\lambda}^{(0)};\widetilde{\boldsymbol{m}},\widetilde{\boldsymbol{\beta}}^*(\cdot))-Q_{2n}(\boldsymbol{\lambda}_0^{(1)},\boldsymbol{\lambda}^{(0)};\widetilde{\boldsymbol{m}},\widetilde{\boldsymbol{\beta}}^*(\cdot))\\
\le& -U_1^\top(\boldsymbol{\lambda}_0^{(1)},\boldsymbol{\lambda}^{(0)};\widetilde{\boldsymbol{m}},\widetilde{\boldsymbol{\beta}}^*(\cdot))\frac{\boldsymbol{u}}{\sqrt{n}}+\frac{1}{n}\boldsymbol{u}^\top\left[-H_{11}(\boldsymbol{\lambda}^{(1)*},\boldsymbol{\lambda}^{(0)};\widetilde{\boldsymbol{m}},\widetilde{\boldsymbol{\beta}}^*(\cdot))\right]\boldsymbol{u}\\
&+\frac{\max(\theta_1,\theta_2)}{\sqrt{n}}\|\boldsymbol{u}\|_1,
\end{aligned}\tag{2.15}
$$

where $\boldsymbol{\lambda}^{(1)*}$ lies in between $\boldsymbol{\lambda}_0^{(1)}$ and $\boldsymbol{\lambda}_0^{(1)}+n^{-1/2}\boldsymbol{u}$ and $\boldsymbol{\lambda}^{(1)*}\xrightarrow{p}\boldsymbol{\lambda}_0^{(1)}$. Here

$$
\frac{1}{\sqrt{n}}U_1(\boldsymbol{\lambda}_0^{(1)},\boldsymbol{\lambda}^{(0)};\widetilde{\boldsymbol{m}},\widetilde{\boldsymbol{\beta}}^*(\cdot))=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^\tau P_1\widetilde{A}_s\left[\boldsymbol{Z}_i-\boldsymbol{E}\left(\bar{a}(s),s\right)\right]dN_i(s), \tag{2.16}
$$

where $\bar{a}(s) \equiv a\left(s, \boldsymbol{\lambda}_0^{(1)}, \boldsymbol{\lambda}^{(0)}; \widetilde{\boldsymbol{m}}, \widetilde{\boldsymbol{\beta}}^*(\cdot)\right)$. Define $B_1 = \mathrm{diag}\left(P_3\left(\boldsymbol{\lambda}_0^{(1)\mathsf{T}}, \boldsymbol{\lambda}^{(0)\mathsf{T}}\right)^{\mathsf{T}}\right)$ and $B_2 = \mathrm{diag}\left(P_4\left(\boldsymbol{\lambda}_0^{(1)\mathsf{T}}, \boldsymbol{\lambda}^{(0)\mathsf{T}}\right)^{\mathsf{T}}\right)$, where P_3 and P_4 are defined in (2.9) and (2.10), respectively. The first-order Taylor expansion of (2.16) around m_0 and $\boldsymbol{\beta}_0^*(s)$ yields

$$
\begin{aligned}
&\frac{1}{\sqrt{n}} U_1(\boldsymbol{\lambda}_0^{(1)}, \boldsymbol{\lambda}^{(0)}; \widetilde{\boldsymbol{m}}, \widetilde{\boldsymbol{\beta}}^*(\cdot)) \\
=&\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \int_0^{\tau} P_1 \widetilde{A}(s) \left[\boldsymbol{Z}_i - \boldsymbol{E}\left(\boldsymbol{\beta}_0(s), s\right)\right] dN_i(s) \\
&- \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \int_0^{\tau} P_1 \widetilde{A}(s) V\left(a^*(s), s\right) B_1 \left(\widetilde{\boldsymbol{m}} - \boldsymbol{m}_0\right) dN_i(s) \\
&- \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \int_0^{\tau} P_1 \widetilde{A}(s) V\left(a^*(s), s\right) B_2 \left(\widetilde{\boldsymbol{\beta}}^*(s) - \boldsymbol{\beta}_0^*(s)\right) dN_i(s),
\end{aligned}
$$

where $a^*(s) = a\left(s, \boldsymbol{\lambda}_0^{(1)}, \boldsymbol{\lambda}^{(0)}; \overline{\boldsymbol{m}}, \overline{\boldsymbol{\beta}}^*(\cdot)\right)$, $\overline{\boldsymbol{m}}$ lies between $\boldsymbol{m}_0$ and $\widetilde{\boldsymbol{m}}$, $\overline{\boldsymbol{\beta}}^*(s)$ lies between $\boldsymbol{\beta}_0^*(s)$ and $\widetilde{\boldsymbol{\beta}}^*(s)$, and $\boldsymbol{E}(\cdot,\cdot)$, $V(\cdot,\cdot)$ are defined in Section 2.3.1. Furthermore, we can prove (a)

$$
\sup_{t\in[0,\tau], \boldsymbol{\beta}\in\mathcal{B}} \|S^{(r)}(\boldsymbol{\beta}, t) - s^{(r)}(\boldsymbol{\beta}, t)\| = O_p(n^{-1/2})
$$

where $\mathcal{B}$ is a compact set of $\mathbb{R}^p$ that includes a neighborhood of $\boldsymbol{\beta}_0(t)$ for $t \in [0, \tau]$; (b) $\widetilde{\boldsymbol{\beta}}(t) \xrightarrow{p} \boldsymbol{\beta}_0(t)$ uniformly in the sense that $\sup_{t\in[0,\tau]} \|\widetilde{\boldsymbol{\beta}}(t) - \boldsymbol{\beta}_0(t)\| = o_p(1)$; (c) both $\boldsymbol{\lambda}^{(1)}$ and $\boldsymbol{\lambda}^{(0)}$ are bounded. (a) can be justified using the central limit theorem for Banach space (Ledoux and Talagrand, 1991). (b) is proved in Appendix A of Tian et al. (2005). (c) can be justified with the following observations: When $\theta_1 = \theta_2 = 0$ and sample size n goes to infinity, approximately the minimizer of (2.3) are $\boldsymbol{\lambda}_1 = \mathbf{1}$ and $\boldsymbol{\lambda}_2 = \mathbf{1}$. Thus, for all $\theta_1 \geq 0$ and $\theta_2 \geq 0$, $\|\boldsymbol{\lambda}_1\|_1 + \|\boldsymbol{\lambda}_2\|_1 < 2p+1$ as $n \to \infty$, and $\boldsymbol{\lambda}^{(1)}$ and $\boldsymbol{\lambda}^{(0)}$ are bounded. Based on (b) and (c), we can prove that $a^*(s) \xrightarrow{p} \boldsymbol{\beta}_0(s)$

uniformly. Thus, given (a), we have

$$\begin{aligned}
&\frac{1}{\sqrt{n}}U_1(\boldsymbol{\lambda}_0^{(1)},\boldsymbol{\lambda}^{(0)};\widetilde{\boldsymbol{m}},\widetilde{\boldsymbol{\beta}}^*(\cdot))\\
\xrightarrow{p}&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int\limits_0^{\tau} P_1A_0(s)\left[\boldsymbol{Z}_i-\boldsymbol{E}\left(\boldsymbol{\beta}_0(s),s\right)\right]dN_i(s)\\
&-\frac{1}{n}\sum_{i=1}^{n}\int\limits_0^{\tau} P_1A_0(s)\left[\frac{Q_2(s)}{Q_0(s)}-\left(\frac{Q_1(s)}{Q_0(s)}\right)^{\otimes 2}\right]dN_i(s)B_1\sqrt{n}\left(\widetilde{\boldsymbol{m}}-\boldsymbol{m}_0\right)\\
&-\frac{1}{n}\sum_{i=1}^{n}\int\limits_0^{\tau} P_1A_0(s)\left[\frac{Q_2(s)}{Q_0(s)}-\left(\frac{Q_1(s)}{Q_0(s)}\right)^{\otimes 2}\right]B_2\sqrt{n}\left(\widetilde{\boldsymbol{\beta}}^*(s)-\boldsymbol{\beta}_0^*(s)\right)dN_i(s)\\
\triangleq& R_1-R_2-R_3,
\end{aligned}$$

where $Q_0(s)$, $Q_1(s)$ and $Q_2(s)$ are defined in Section 2.3.1.

Next we want to show $R_1-R_2-R_3=O_p(1)$. Using the fact that

$$Q_2(s)/Q_0(s)-(Q_1(s)/Q_0(s))^{\otimes 2}=O_p(1),$$

$A_0(s)=O_p(1)$ and $\sqrt{n}\left(\widetilde{\boldsymbol{m}}-\boldsymbol{m}_0\right)=O_p(1)$, it is easy to verify that $R_2=O_p(1)$. Further, similarly as Lemma 2.2 we can show

$$\begin{aligned}
R_3\approx&\frac{1}{n}\sum_{i=1}^{n}\int\limits_0^{\tau}(h_n)^{-1/2}P_1A_0(s)\left[\frac{Q_2(s)}{Q_0(s)}-\left(\frac{Q_1(s)}{Q_0(s)}\right)^{\otimes 2}\right]B_2\Sigma^{-1}(s)\\
&(nh_n)^{-1/2}\sum_{j=1}^{n}\int_0^{\tau}\left(\boldsymbol{Z}_j(u)-E\left(\boldsymbol{\beta}_0(s),u\right)\right)K\left(\frac{u-s}{h_n}\right)dM_j(u)dN_i(s)\\
\xrightarrow{p}&\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\int_0^{\tau}\left\{\int_0^{\tau}P_1A_0(s)\left[\frac{Q_2(s)}{Q_0(s)}-\left(\frac{Q_1(s)}{Q_0(s)}\right)^{\otimes 2}\right]\right.\\
&\left.B_2\Sigma^{-1}(s)\left(\boldsymbol{Z}_j(u)-E\left(\boldsymbol{\beta}_0(s),u\right)\right)\frac{1}{h_n}K\left(\frac{u-s}{h_n}\right)Q_0(s)ds\right\}dM_j(u)\\
=&\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\int_0^{\tau}P_1A_0(u)\Sigma(u)B_2\Sigma^{-1}(u)\left(\boldsymbol{Z}_j(u)-E\left(\boldsymbol{\beta}_0(u),u\right)\right)dM_j(u),
\end{aligned}$$

where the "$\xrightarrow{p}$" sign can be proved using the strong approximation argument (Yandell, 1983). By the martingale central limit theorem (Andersen and Gill, 1982), $R_3 = O_p(1)$. Then, we only need to prove $R_1 = O_p(1)$. Since

$$R_1 = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \int_0^{\tau} P_1 A_0(s) \left[\boldsymbol{Z}_i - \boldsymbol{E}\left(\boldsymbol{\beta}_0(s), s\right)\right] dM_i(s) + o_p(1),$$

it follows from the martingale central limit theorem (Andersen and Gill, 1982) that $R_1 = O_p(1)$. Thus,

$$\frac{1}{\sqrt{n}} U_1(\boldsymbol{\lambda}_0^{(1)}, \boldsymbol{\lambda}^{(0)}; \widetilde{\boldsymbol{m}}, \widetilde{\boldsymbol{\beta}}^*(\cdot)) \xrightarrow{p} R_1 - R_2 - R_3 = O_p(1).$$

Furthermore, we know

$$-\frac{1}{n} H_{11}(\boldsymbol{\lambda}^{(1)*}, \boldsymbol{\lambda}^{(0)}; \widetilde{\boldsymbol{m}}, \widetilde{\boldsymbol{\beta}}^*(\cdot)) = \frac{1}{n} \sum_{i=1}^{n} \int_0^{\tau} P_1 \widetilde{A}(s) V(\ddot{a}(s), s) \left(P_1 \widetilde{A}(s)\right)^{\intercal} dN_i(s)$$

$$\xrightarrow{p} \int_0^{\tau} P_1 A_0(s) \Sigma(s) \left(P_1 A_0(s)\right)^{\intercal} ds \triangleq I_{11},$$

where $\ddot{a}(s) = a(s, \boldsymbol{\lambda}^{(1)*}, \boldsymbol{\lambda}^{(0)}; \widetilde{\boldsymbol{m}}, \widetilde{\boldsymbol{\beta}}^*(\cdot))$. From the definition of P_1 (2.7) and $A_0(s)$ (2.13), we can show that I_{11} is positive definite and $I_{11} = O_p(1)$. Therefore, given $\max(\theta_1, \theta_2)/\sqrt{n}$ is bounded, if we choose a sufficient large C, the second term in (2.15) is of order C^2. The first and third terms are of order C, which are dominated by the second term. Therefore (2.14) holds and it completes the proof.

2.6.3 Proof of Theorem 2.2

For any $\boldsymbol{\lambda}^{(1)}$ satisfy $\|\boldsymbol{\lambda}^{(1)} - \boldsymbol{\lambda}_0^{(1)}\| = O_p(n^{-1/2})$, we have

$$\begin{aligned}
&Q_{2n}(\boldsymbol{\lambda}^{(1)}, \boldsymbol{\lambda}^{(0)}; \widetilde{\boldsymbol{m}}, \widetilde{\boldsymbol{\beta}}(\cdot)) - Q_{2n}(\boldsymbol{\lambda}^{(1)}, \mathbf{0}; \widetilde{\boldsymbol{m}}, \widetilde{\boldsymbol{\beta}}(\cdot)) \\
\geq& - U_2^{\mathsf{T}}(\boldsymbol{\lambda}^{(1)}, \boldsymbol{\lambda}^{(0)*}; \widetilde{\boldsymbol{m}}, \widetilde{\boldsymbol{\beta}}(\cdot))\boldsymbol{\lambda}^{(0)} + \min(\theta_1, \theta_2) \|\boldsymbol{\lambda}^{(0)}\|_1 \\
\geq& \left[- \sup_k \left| U_{2k}(\boldsymbol{\lambda}^{(1)}, \boldsymbol{\lambda}^{(0)*}; \widetilde{\boldsymbol{m}}, \widetilde{\boldsymbol{\beta}}(\cdot)) \right| + \min(\theta_1, \theta_2) \right] \|\boldsymbol{\lambda}^{(0)}\|_1,
\end{aligned} \tag{2.17}$$

where $U_{2k}(\boldsymbol{\lambda}^{(1)}, \boldsymbol{\lambda}^{(0)*}; \widetilde{\boldsymbol{m}}, \widetilde{\boldsymbol{\beta}}(\cdot))$ is the kth component of the vector $U_2(\boldsymbol{\lambda}^{(1)}, \boldsymbol{\lambda}^{(0)*}; \widetilde{\boldsymbol{m}}, \widetilde{\boldsymbol{\beta}}(\cdot))$ and $\boldsymbol{\lambda}^{(0)*}$ lies in between $\boldsymbol{\lambda}^{(0)}$ and $\mathbf{0}$. Define $B_3 = \operatorname{diag}\left(P_3\left(\boldsymbol{\lambda}^{(1)^{\mathsf{T}}}, \boldsymbol{\lambda}^{(0)*^{\mathsf{T}}}\right)^{\mathsf{T}}\right)$ and $B_4 = \operatorname{diag}\left(P_4\left(\boldsymbol{\lambda}^{(1)^{\mathsf{T}}}, \boldsymbol{\lambda}^{(0)*^{\mathsf{T}}}\right)^{\mathsf{T}}\right)$. The first order Taylor expansion of $U_2(\boldsymbol{\lambda}^{(1)}, \boldsymbol{\lambda}^{(0)*}; \widetilde{\boldsymbol{m}}, \widetilde{\boldsymbol{\beta}}(\cdot))$ around m_0 and $\boldsymbol{\beta}_0^*(s)$ yields

$$\begin{aligned}
&\frac{\sqrt{nh_n}}{\sqrt{n}} U_2(\boldsymbol{\lambda}^{(1)}, \boldsymbol{\lambda}^{(0)*}; \widetilde{\boldsymbol{m}}, \widetilde{\boldsymbol{\beta}}^*(\cdot)) \triangleq \frac{\sqrt{nh_n}}{\sqrt{n}} \sum_{i=1}^n \int_0^\tau P_2 \widetilde{A}(s) \left[\boldsymbol{Z}_i - \boldsymbol{E}(a(s), s)\right] dN_i(s) \\
=& \frac{\sqrt{nh_n}}{\sqrt{n}} \sum_{i=1}^n \int_0^\tau P_2 \widetilde{A}(s) \left[\boldsymbol{Z}_i - \boldsymbol{E}(a^*(s), s)\right] dN_i(s) \\
&- \frac{\sqrt{nh_n}}{\sqrt{n}} \sum_{i=1}^n \int_0^\tau P_2 \widetilde{A}(s) V(\overline{a}(s), s) B_3 \left(\widetilde{\boldsymbol{m}} - \boldsymbol{m}_0\right) dN_i(s) \\
&- \frac{\sqrt{nh_n}}{\sqrt{n}} \sum_{i=1}^n \int_0^\tau P_1 \widetilde{A}(s) V(\overline{a}(s), s) B_4 \left(\widetilde{\boldsymbol{\beta}}^*(s) - \boldsymbol{\beta}_0^*(s)\right) dN_i(s) \\
\xrightarrow{p}& \frac{1}{\sqrt{n}} \sum_{i=1}^n \int_0^\tau \sqrt{nh_n} P_2 \widetilde{A}(s) \left[\boldsymbol{Z}_i - \boldsymbol{E}(a^*(s), s)\right] dN_i(s) \\
&- \frac{1}{n} \sum_{i=1}^n \int_0^\tau \sqrt{nh_n} P_2 \widetilde{A}(s) \left[\frac{Q_2(s)}{Q_0(s)} - \left(\frac{Q_1(s)}{Q_0(s)}\right)^{\otimes 2}\right] dN_i(s) B_3 \sqrt{n} \left(\widetilde{\boldsymbol{m}} - \boldsymbol{m}_0\right) \\
&- \frac{1}{n} \sum_{i=1}^n \int_0^\tau \sqrt{nh_n} P_2 \widetilde{A}(s) \left[\frac{Q_2(s)}{Q_0(s)} - \left(\frac{Q_1(s)}{Q_0(s)}\right)^{\otimes 2}\right] B_4 \sqrt{n} \left(\widetilde{\boldsymbol{\beta}}^*(s) - \boldsymbol{\beta}_0^*(s)\right) dN_i(s) \\
\triangleq& R_4 - R_5 - R_6,
\end{aligned}$$

where $a(s) = a\left(s, \boldsymbol{\lambda}^{(1)}, \boldsymbol{\lambda}^{(0)*}; \widetilde{\boldsymbol{m}}, \widetilde{\boldsymbol{\beta}}^*(\cdot)\right)$, $a^*(s) = a\left(s, \boldsymbol{\lambda}^{(1)}, \boldsymbol{\lambda}^{(0)*}; \boldsymbol{m}_0, \boldsymbol{\beta}_0^*(\cdot)\right)$, $\overline{a}(s) = a\left(s, \boldsymbol{\lambda}^{(1)}, \boldsymbol{\lambda}^{(0)*}; \overline{\boldsymbol{m}}, \overline{\boldsymbol{\beta}}^*(\cdot)\right)$ and $\boldsymbol{E}(\cdot,\cdot)$, $Q_0(\cdot)$, $Q_1(\cdot)$, $Q_2(\cdot)$ are defined in Section 2.3.1. Here $\overline{\boldsymbol{m}}$ lies in between $\boldsymbol{m}_0$ and $\widetilde{\boldsymbol{m}}$, and $\overline{\boldsymbol{\beta}}^*(s)$ lies in between $\boldsymbol{\beta}_0^*(s)$ and $\widetilde{\boldsymbol{\beta}}^*(s)$. Using the fact that $\sqrt{nh_n}P_2\widetilde{A}(s) = O_p(1)$, $Q_2(s)/Q_0(s) - (Q_1(s)/Q_0(s))^{\otimes 2} = O_p(1)$ and Lemma 2.1 and 2.2, it is easy to verify $R_5 = O_p(1)$ and $R_6 = O_p(1)$. For R_4, by the first order Taylor expansion of $\boldsymbol{\lambda}^{(1)}$ around $\boldsymbol{\lambda}_0^{(1)}$, we have

$$\begin{aligned}
R_4 \equiv & \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^\tau \sqrt{nh_n}P_2\widetilde{A}(s)\left[\boldsymbol{Z}_i - \boldsymbol{E}\left(a^*(s), s\right)\right]dN_i(s) \\
= & \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^\tau \sqrt{nh_n}P_2\widetilde{A}(s)\left[\boldsymbol{Z}_i - \boldsymbol{E}\left(\boldsymbol{\beta}_0(s), s\right)\right]dN_i(s) \\
& - \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \sqrt{nh_n}P_2\widetilde{A}(s)V(\ddot{a}(s), s)B_5(s)dN_i(s)\sqrt{n}\left(\lambda^{(1)} - \lambda_0^{(1)}\right) \\
= & O_p(1) - O_p(1) = O_p(1),
\end{aligned}$$

where $\ddot{a}(s) = a(s, \boldsymbol{\lambda}^{(1)*}, \boldsymbol{\lambda}^{(0)*}; \boldsymbol{m}_0, \boldsymbol{\beta}_0^*(\cdot))$, $B_5(s) = \mathrm{diag}(\boldsymbol{m}_0)P_{31} + \mathrm{diag}(\boldsymbol{\beta}_0^*(s))P_{41}$, $\boldsymbol{\lambda}^{(1)*}$ lies in between $\boldsymbol{\lambda}^{(1)}$ and $\boldsymbol{\lambda}_0^{(1)}$, and P_{31}, P_{41} are defined in (2.9) and (2.10), respectively. Thus $\sqrt{h_n}U_2(\boldsymbol{\lambda}^{(1)}, \boldsymbol{\lambda}^{(0)*}; \widetilde{\boldsymbol{m}}, \widetilde{\boldsymbol{\beta}}^*(\cdot)) = O_p(1)$. From (2.17),

$$\begin{aligned}
& Q_{2n}(\boldsymbol{\lambda}^{(1)}, \boldsymbol{\lambda}^{(0)}; \widetilde{\boldsymbol{m}}, \widetilde{\boldsymbol{\beta}}(\cdot)) - Q_{2n}(\boldsymbol{\lambda}^{(1)}, 0; \widetilde{\boldsymbol{m}}, \widetilde{\boldsymbol{\beta}}(\cdot)) \\
\geq & \left[-O_p(h_n^{-1/2}) + \min(\theta_1, \theta_2)\right]\|\boldsymbol{\lambda}^{(0)}\|_1.
\end{aligned}$$

If $h_n^{1/2}\min(\theta_1, \theta_2) \to \infty$, then with probability one

$$Q_{2n}(\boldsymbol{\lambda}^{(1)}, \boldsymbol{\lambda}^{(0)}; \widetilde{\boldsymbol{m}}, \widetilde{\boldsymbol{\beta}}^*(s)) - Q_{2n}(\boldsymbol{\lambda}^{(1)}, \boldsymbol{0}; \widetilde{\boldsymbol{m}}, \widetilde{\boldsymbol{\beta}}^*(s)) > 0 \text{ for all } \boldsymbol{\lambda}^{(0)} \neq 0.$$

It completes the proof.

2.6.4 Proof of Theorem 2.3

(a) From Theorm 2.1, $\|\widehat{\boldsymbol{\lambda}}^{(1)} - \boldsymbol{\lambda}_0^{(1)}\| = O_p(n^{-1/2})$ where $\boldsymbol{\lambda}_0^{(1)} = \mathbf{1}$. Thus $P(\widehat{\boldsymbol{\lambda}}_k^{(1)} \neq 0) \to 1$ for all k. This in couple with $P(\widehat{\boldsymbol{\lambda}}^{(0)} = \mathbf{0}) \to 1$ from Theorem 2.2 proves $\widehat{\mathrm{I}}_O = \mathrm{I}_O$, $\widehat{\mathrm{I}}_C = \mathrm{I}_C$ and $\widehat{\mathrm{I}}_{NC} = \mathrm{I}_{NC}$ hold with probability tending to one. It completes the proof for part (a).

(b) From part (a), we know with probability tending to one, $\widehat{\boldsymbol{\beta}}_C(t) \equiv \widehat{\boldsymbol{\beta}}_C$. Thus,

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_C - \boldsymbol{\beta}_C\right) = \sqrt{n}\left(\widetilde{\boldsymbol{m}}_C \circ \widehat{\boldsymbol{\lambda}}_1^C - \boldsymbol{m}_C\right) = \widetilde{\boldsymbol{m}}_C \circ \sqrt{n}\left(\widehat{\boldsymbol{\lambda}}_1^C - \boldsymbol{\lambda}_{01}^C\right) + \sqrt{n}\left(\widetilde{\boldsymbol{m}}_C - \boldsymbol{m}_C\right),$$

where $\widetilde{\boldsymbol{m}}_C$, $\widehat{\boldsymbol{\lambda}}_1^C$, $\boldsymbol{m}_C$ are sub-vectors of $\widetilde{\boldsymbol{m}}$, $\widehat{\boldsymbol{\lambda}}_1$, $\boldsymbol{m}_0$ with indexes in I_C. Since $\widehat{\boldsymbol{\lambda}}_1^C$ and $\widetilde{\boldsymbol{m}}_C$ are root-n consistent estimator for $\boldsymbol{\lambda}_{01}^C$ and $\boldsymbol{m}_C$ respectively. We have $\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_C - \boldsymbol{\beta}_C\right) = O_p(1)$. It completes the proof for part (b).

(c) From the proofs of part (a), and Theorem 2.1 and 2.2, it is easy to show that $\widehat{\boldsymbol{\lambda}}^{(1)}$ is a root-n consistent maximizer of $Q_{2n}(\boldsymbol{\lambda}^{(1)}, \mathbf{0}; \widetilde{\boldsymbol{m}}, \widetilde{\boldsymbol{\beta}}^*(\cdot))$. Thus

$$\left.\frac{\partial Q_{2n}(\boldsymbol{\lambda}^{(1)}, \mathbf{0}; \widetilde{\boldsymbol{m}}, \widetilde{\boldsymbol{\beta}}^*(\cdot))}{\partial \boldsymbol{\lambda}^{(1)}}\right|_{\boldsymbol{\lambda}^{(1)} = \widehat{\boldsymbol{\lambda}}^{(1)}} = \mathbf{0}.$$

This is equivalent to

$$-\frac{1}{\sqrt{n}} U_1(\widehat{\boldsymbol{\lambda}}^{(1)}, \mathbf{0}; \widetilde{\boldsymbol{m}}, \widetilde{\boldsymbol{\beta}}^*(\cdot)) + \frac{\overline{\boldsymbol{\theta}}}{\sqrt{n}} = \mathbf{0}, \tag{2.18}$$

where $\overline{\boldsymbol{\theta}}$ is a vector of length $(p_2 + 2p_3)$ consisting with elements θ_1 and θ_2.

The first order Taylor expansion of $n^{-1/2}U_1(\widehat{\boldsymbol{\lambda}}^{(1)}, \mathbf{0}; \widetilde{\boldsymbol{m}}, \widetilde{\boldsymbol{\beta}}^*(\cdot))$ around $\boldsymbol{m}_0$ and $\boldsymbol{\beta}_0^*(\cdot)$ yields

$$
\begin{aligned}
&\frac{1}{\sqrt{n}}U_1(\widehat{\boldsymbol{\lambda}}^{(1)}, \mathbf{0}; \widetilde{\boldsymbol{m}}, \widetilde{\boldsymbol{\beta}}^*(\cdot)) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^{\tau} P_1\widetilde{A}(s)\left[\boldsymbol{Z}_i - \boldsymbol{E}\left(\widetilde{a}(s), s\right)\right] dN_i(s) \\
&- \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^{\tau} P_1\widetilde{A}(s)V\left(a^*(s), s\right)\widetilde{B}_6\left(\widetilde{\boldsymbol{m}} - \boldsymbol{m}_0\right) dN_i(s) \\
&- \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^{\tau} P_1\widetilde{A}(s)V\left(a^*(s), s\right)\widetilde{B}_7\left(\widetilde{\boldsymbol{\beta}}^*(s) - \boldsymbol{\beta}_0^*(s)\right) dN_i(s) \\
\triangleq\, & R_7 - R_8 - R_9,
\end{aligned}
\tag{2.19}
$$

where $\widetilde{a}(s) = a(s, \widehat{\boldsymbol{\lambda}}^{(1)}, \mathbf{0}; \boldsymbol{m}_0, \boldsymbol{\beta}_0^*(\cdot))$, $a^*(s) = a(s, \widehat{\boldsymbol{\lambda}}^{(1)}, \mathbf{0}; \overline{\boldsymbol{m}}, \overline{\boldsymbol{\beta}}^*(\cdot))$,

$$
\widetilde{B}_6 = \operatorname{diag}\left(P_3\left(\left(\widehat{\boldsymbol{\lambda}}^{(1)}\right)^{\intercal}, \mathbf{0}^{\intercal}\right)^{\intercal}\right) \text{ and } \widetilde{B}_7 = \operatorname{diag}\left(P_4\left(\left(\widehat{\boldsymbol{\lambda}}^{(1)}\right)^{\intercal}, \mathbf{0}^{\intercal}\right)^{\intercal}\right).
$$

Using similar arguments for the proofs of Theorem 2.1 and 2.2, we can show

$$
\begin{aligned}
R_8 &\xrightarrow{p} \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau} P_1A_0(s)\left[\frac{Q_2(s)}{Q_0(s)} - \left(\frac{Q_1(s)}{Q_0(s)}\right)^{\otimes 2}\right] dN_i(s)B_6\sqrt{n}\left(\widetilde{\boldsymbol{m}} - \boldsymbol{m}_0\right) \\
&\xrightarrow{p} \frac{1}{\tau}\int_0^{\tau} P_1A_0(s)\Sigma(s)dsB_6\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^{\tau}\Sigma^{-1}(u)\left(\boldsymbol{Z}_i(u) - \boldsymbol{E}\left(\boldsymbol{\beta}_0(u), u\right)\right) dM_i(u)\right],
\end{aligned}
\tag{2.20}
$$

and

$$
\begin{aligned}
R_9 \xrightarrow{p} & \frac{1}{nh_n^{1/2}} \sum_{i=1}^{n} \int\limits_0^{\tau} P_1 A_0(s) \left[\frac{Q_2(s)}{Q_0(s)} - \left(\frac{Q_1(s)}{Q_0(s)} \right)^{\otimes 2} \right] B_7 \sqrt{nh_n} \left(\widetilde{\boldsymbol{\beta}}^*(s) - \boldsymbol{\beta}_0^*(s) \right) dN_i(s) \\
\approx & \frac{1}{nh_n^{1/2}} \sum_{i=1}^{n} \int\limits_0^{\tau} P_1 A_0(s) \left[\frac{Q_2(s)}{Q_0(s)} - \left(\frac{Q_1(s)}{Q_0(s)} \right)^{\otimes 2} \right] B_7 \\
& \Sigma^{-1}(s)(nh_n)^{-1/2} \sum_{j=1}^{n} \int_0^{\tau} \left(\boldsymbol{Z}_j(u) - \boldsymbol{E}\left(\boldsymbol{\beta}_0(s), u \right) \right) K\left(\frac{u-s}{h_n} \right) dM_j(u) dN_i(s) \\
= & \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \int_0^{\tau} \left\{ \frac{1}{n} \sum_{i=1}^{n} \int_0^{\tau} P_1 A_0(s) \left[\frac{Q_2(s)}{Q_0(s)} - \left(\frac{Q_1(s)}{Q_0(s)} \right)^{\otimes 2} \right] B_7 \Sigma^{-1}(s) \right. \\
& \left. \left(\boldsymbol{Z}_j(u) - \boldsymbol{E}\left(\boldsymbol{\beta}_0(s), u \right) \right) \frac{1}{h_n} K\left(\frac{u-s}{h_n} \right) dN_i(s) \right\} dM_j(u) \\
\xrightarrow{p} & \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \int_0^{\tau} P_1 A_0(u) \Sigma(u) B_7 \Sigma^{-1}(u) \left(\boldsymbol{Z}_j(u) - \boldsymbol{E}\left(\boldsymbol{\beta}_0(u), u \right) \right) dM_j(u), \qquad (2.21)
\end{aligned}
$$

where

$$
B_6 = \begin{pmatrix} \mathbf{0}_{p_1 \times p_1} & \mathbf{0}_{p_1 \times (p_2+p_3)} \\ \mathbf{0}_{(p_2+p_3) \times p_1} & \boldsymbol{I}_{p_2+p_3} \end{pmatrix}; B_7 = \begin{pmatrix} \mathbf{0}_{(p_1+p_2) \times (p_1+p_2)} & \mathbf{0}_{(p_1+p_2) \times p_3} \\ \mathbf{0}_{p_3 \times (p_1+p_2)} & \boldsymbol{I}_{p_3} \end{pmatrix}.
$$

Here B_6 and B_7 are the limits of $\widetilde{B}_6$ and $\widetilde{B}_7$, respectively. The first order Taylor expansion of R_7 around $\boldsymbol{\lambda}_0^{(1)}$ yields

$$
\begin{aligned}
R_7 = & \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \int_0^{\tau} P_1 \widetilde{A}(s) \left[\boldsymbol{Z}_i - \boldsymbol{E}\left(\boldsymbol{\beta}_0(s), s \right) \right] dN_i(s) \\
& - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \int_0^{\tau} P_1 \widetilde{A}(s) V\left(\ddot{a}(s), s \right) B_8(s) dN_i(s) \sqrt{n} (\widehat{\boldsymbol{\lambda}}^{(1)} - \boldsymbol{\lambda}_0^{(1)}) \\
\xrightarrow{p} & \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \int_0^{\tau} P_1 A_0(s) \left[\boldsymbol{Z}_i - \boldsymbol{E}\left(\boldsymbol{\beta}_0(s), s \right) \right] dM_i(s) - F \sqrt{n} (\widehat{\boldsymbol{\lambda}}^{(1)} - \boldsymbol{\lambda}_0^{(1)}), \qquad (2.22)
\end{aligned}
$$

where

$$B_8(s) = \begin{pmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \operatorname{diag}(\boldsymbol{m}_C) & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \operatorname{diag}(\boldsymbol{m}_{NC}) & \operatorname{diag}(\boldsymbol{\beta}^*_{NC}(s)) \end{pmatrix}$$

and

$$F = \int_0^\tau P_1 A_0(s)\Sigma(s)B_8(s)ds.$$

Here $\ddot{a}(s) = a(s, \boldsymbol{\lambda}^{(1)*}, \mathbf{0}; \boldsymbol{m}_0, \boldsymbol{\beta}^*_0(\cdot))$ and $\boldsymbol{\lambda}^{(1)*}$ lies in between $\boldsymbol{\lambda}_0^{(1)}$ and $\widehat{\boldsymbol{\lambda}}^{(1)}$.

Under the assumption $\max(\theta_1, \theta_2)/\sqrt{n} \to 0$, combining (2.18), (2.19), (2.20), (2.21), (2.22), we have

$$\sqrt{n}\left(\widehat{\boldsymbol{\lambda}}^{(1)} - \boldsymbol{\lambda}_0^{(1)}\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \int_0^\tau F^{-1}G(u)\left[\boldsymbol{Z}_i - \boldsymbol{E}\left(\boldsymbol{\beta}_0(s), s\right)\right]dM_i(s) + o_p(1), \tag{2.23}$$

where

$$G(u) = P_1A_0(u) - \frac{1}{\tau}\left(\int_0^\tau P_1A_0(s)\Sigma(s)ds\right)B_6\Sigma^{-1}(u) - P_1A_0(u)\Sigma(u)B_7\Sigma^{-1}(u).$$

Because

$$\begin{aligned}
&\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_C - \boldsymbol{\beta}_C\right) = \widetilde{\boldsymbol{m}}_C \circ \sqrt{n}\left(\widehat{\boldsymbol{\lambda}}_1^C - \boldsymbol{\lambda}_{01}^C\right) + n^{1/2}\left(\widetilde{\boldsymbol{m}}_C - \boldsymbol{m}_C\right) \\
=&\operatorname{diag}(\boldsymbol{m}_C)\, B_9\sqrt{n}\left(\widehat{\boldsymbol{\lambda}}^{(1)} - \boldsymbol{\lambda}_0^{(1)}\right) \\
&+ \frac{1}{\sqrt{n}}\sum_{i=1}^n\int_0^\tau \frac{1}{\tau}B_{10}\Sigma^{-1}(u)\left[\boldsymbol{Z}_i - \boldsymbol{E}\left(\boldsymbol{\beta}_0(u), u\right)\right]dM_i(u) + o_p(1),
\end{aligned} \tag{2.24}$$

where $B_9 = (\boldsymbol{I}_{p_2}|\mathbf{0}_{p_2\times p_3}|\mathbf{0}_{p_2\times p_3})$ and $B_{10} = (\mathbf{0}_{p_2\times p_1}|\boldsymbol{I}_{p_2}|\mathbf{0}_{p_2\times p_3})$. Combining (2.23) and (2.24), we have

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_C - \boldsymbol{\beta}_C\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^n\int_0^\tau\left[D(u) + \frac{1}{\tau}B_{10}\Sigma^{-1}(u)\right]\left(\boldsymbol{Z}_i - \boldsymbol{E}\left(\boldsymbol{\beta}_0(u), u\right)\right)dM_i(u) + o_p(1),$$

where

$$D(u) = \operatorname{diag}(\boldsymbol{m}_C)\, B_9 F^{-1} G(u). \tag{2.25}$$

Finally by the martingale central limit theorem (Andersen and Gill, 1982), $n^{1/2}\left(\widehat{\boldsymbol{\beta}}_C - \boldsymbol{\beta}_C\right)$ converges weakly to a normal distribution with mean 0 and variance Σ_m^F, where

$$\Sigma_m^F = \int_0^\tau \left(D(u) + \frac{1}{\tau} B_{10}\Sigma^{-1}(u)\right) \Sigma(u) \left(D(u) + \frac{1}{\tau} B_{10}\Sigma^{-1}(u)\right)^\intercal du.$$

(d) Because

$$\begin{aligned}
&(nh_n)^{1/2}\left(\widehat{\boldsymbol{\beta}}_{NC}(t) - \boldsymbol{\beta}_{NC}(t)\right) = (nh_n)^{1/2}\left(\widetilde{\boldsymbol{m}}_{NC} \circ \widehat{\boldsymbol{\lambda}}_1^{NC} + \widetilde{\boldsymbol{\beta}}_{NC}^* \circ \widehat{\boldsymbol{\lambda}}_2^{NC} - \boldsymbol{\beta}_{NC}(t)\right) \\
=&(nh_n)^{1/2}\left[\widetilde{\boldsymbol{m}}_{NC} \circ \left(\widehat{\boldsymbol{\lambda}}_1^{NC} - \boldsymbol{\lambda}_{01}^{NC}\right) + \widetilde{\boldsymbol{\beta}}_{NC}^*(t) \circ \left(\widehat{\boldsymbol{\lambda}}_2^{NC} - \boldsymbol{\lambda}_{02}^{NC}\right) + \left(\widetilde{\boldsymbol{\beta}}_{NC}(t) - \boldsymbol{\beta}_{NC}(t)\right)\right] \\
=&(nh_n)^{1/2}\left(\widetilde{\boldsymbol{\beta}}_{NC}(t) - \boldsymbol{\beta}_{NC}(t)\right) + o_p(1) \\
\xrightarrow{d}& N\left\{0,\ \{\Sigma^{-1}(t)\}_{NC,NC}\int_{-1}^1 K^2(s)ds\right\},
\end{aligned}$$

where $\{\Sigma^{-1}(t)\}_{NC,NC}$ is the submatrix of $\Sigma^{-1}(t)$ corresponding to I_{NC}. It completes the proof for part (d).

Chapter 3

ConvexLAR: An Extension of Least Angle Regression

3.1 Introduction

Regularization is a tool to avoid overfitting and obtain parsimonious and interpretable models, especially when the number of parameters exceeds the number of observations. One powerful regularization technique is penalization. In general, a penalty method minimizes the sum of a loss function and a penalty term. The simplest ℓ_1 penalty leads to the popular lasso regression (Tibshirani, 1996; Donoho and Johnstone, 1994). Various other penalty methods have been developed thereafter. Each one targets on a specific question that arises in applications. For instance, the group penalty (Yuan and Lin, 2006; Meier et al., 2008) aims to select groups of variables such as in factorial data analysis. The adaptive lasso (Zou, 2006) applies weighted penalty method in a data driven fashion that improves the asymptotic properties. All these penalty methods are formulated as an optimization problem and the solution is obtained by either optimizing at a grid of tuning parameter values or by a path algorithm that tracks the solution as a function of the tuning parameter.

In contrast to penalty methods, the least angle regression (LAR) proposed by Efron et al.

(2004) is purely motivated by a geometric argument rather than optimization. Given a response vector $\boldsymbol{y} = (y_1, y_2, \cdots, y_n)^\intercal \in \mathbb{R}^n$ and corresponding design matrix $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^\intercal \in \mathbb{R}^{n\times p}$ with $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \cdots, x_{ip})^\intercal$, let $\mathcal{A}_t$ be the active index set at time t. Following Efron et al. (2004), we assume that the covariates $\boldsymbol{x}^{(j)} = (x_{1j}, x_{2j}, \ldots, x_{nj})^\intercal$ have mean 0 and unit length, and that the response $\boldsymbol{y}$ has mean 0. At $t = 0$, $\boldsymbol{\beta}(0) = \mathbf{0}_p$ and $\mathcal{A}_0$ contains the predictor that is most correlated with the response vector $\boldsymbol{y}$. At any $t > 0$, regression coefficients of active predictors, namely $\beta_j(t)$ for $j \in \mathcal{A}_t$, move along a direction such that their corresponding predictor vectors $\boldsymbol{x}^{(j)}$ share the same absolute correlation (angle) with the residual vector $\boldsymbol{e}(t) = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}(t)$. Here the correlation is nothing but the scaled score vector of the least squares criterion with entries

$$-\frac{1}{2}\frac{\partial}{\partial\beta_j}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^\intercal\boldsymbol{\beta})^2\bigg|_{\boldsymbol{\beta}=\boldsymbol{\beta}(t)}.$$

LAR has gained wide popularity since its introduction. However, there seem only a few attempts for generalizations, strikingly unparallel to the fast development of penalty methods. Specific versions of group LAR for least squares problem are mentioned in Yuan and Lin (2006) and Park and Hastie (2006). Wu extends LAR to the generalized linear models (Wu, 2011) and the Cox's proportional hazard model (Wu, 2012). In this article, we demonstrate that the powerful geometric idea of LAR can be generalized in a fruitful way and leads to potentially many more applications.

The remaining of the chapter is organized as follows. In Section 3.2, we derive a basic ConvexLAR algorithm that performs continuous variable selection for any general convex loss functions. For least squares loss, Efron et al. (2004) show that the LAR solution path is piecewise linear. This leads to their efficient path following algorithm with computational cost of a single ordinary least squares estimation. For a general loss function, the piecewise linearity property is lost. However, it can be shown that the solution path is piecewise smooth and, within each path segment, follows a simple ordinary differential equation (ODE). ConvexLAR tracks the solution path by utilizing the rich numerical resource for solving ODE. Just like the original LAR for least squares, a simple modification of ConvexLAR yields the corresponding lasso solution path.

In Section 3.3, we show that the geometric idea of LAR can be adapted to various situations. We demonstrate this by incorporating data adaptive weights and group selection into the ConvexLAR algorithm. Comparing to their penalization analogs, these extensions avoid repeated optimizations and are computationally attractive. Moreover, a slight modification of ConvexLAR yields the exact solution path for the corresponding penalization method. ConvexLAR and its extensions are illustrated by various numerical examples in Section 3.4. To the best of our knowledge, no LAR algorithms have been proposed for these examples in the current literature. Finally we conclude with a brief discussion. The algorithm is implemented in Matlab and corresponding code can be downloaded from `http://www4.ncsu.edu/~wxiao`.

3.2 ConvexLAR Algorithm and ConvexLASSO Modification

In this section, we derive the ConvexLAR algorithm which forms the basis of various extensions in the next section. The algorithm is similar to the LAR algorithms developed for GLM and Cox model (Wu, 2011, 2012) but with much more generality and simpler derivation. We consider an arbitrary strictly convex loss function $f(\boldsymbol{\beta})$, where $\boldsymbol{\beta} \in \mathrm{I\!R}^p$ is the vector of parameters subject to regularization. $\nabla f(\boldsymbol{\beta}) = [\nabla_1 f(\boldsymbol{\beta}), \ldots, \nabla_p f(\boldsymbol{\beta})]^\mathsf{T} \in \mathrm{I\!R}^p$ denotes the gradient vector of the loss function and $\boldsymbol{H}(\boldsymbol{\beta}) = d^2 f(\boldsymbol{\beta}) \in \mathrm{I\!R}^{p\times p}$ the Hessian. When f is a negative log-likelihood function, $\nabla f(\boldsymbol{\beta})$ equals the negative of the score vector and $\boldsymbol{H}(\boldsymbol{\beta})$ is the observed information matrix. We use t to index the LAR solution path, with the solution at any t denoted by $\boldsymbol{\beta}(t)$. The active index set at t is denoted by $\mathcal{A}_t$. For notational simplicity, we will drop the subscript t whenever it is obvious from the context. For instance, $\boldsymbol{\beta}_{\mathcal{A}}(t)$ is the subvector of $\boldsymbol{\beta}(t)$ and $\boldsymbol{H}_{\mathcal{A}}(\boldsymbol{\beta}(t))$ is the submatrix of the Hessian corresponding to active predictors at t.

The key idea of LAR (Efron et al., 2004) is to move the solution in a direction such that the gradient (score) corresponding to each active predictor variable has the same absolute value. We denote this common value by $s(t)$, where s stands for the score. This prescribes that the active solution vector has to satisfy $|\nabla_j f(\boldsymbol{\beta})| = \operatorname{sgn}(\nabla_j f(\boldsymbol{\beta})) \cdot \nabla_j f(\boldsymbol{\beta}) = s(t)$ or equivalently

$$\nabla_j f(\boldsymbol{\beta}) - \operatorname{sgn}(\nabla_j f(\boldsymbol{\beta}))s(t) = 0, \quad j \in \mathcal{A}_t.$$

Note that, for any active predictor j, $\nabla_j f(\boldsymbol{\beta})$ is non-zero and therefore has a constant sign, denoted by $\operatorname{sgn}(\nabla_j f(\boldsymbol{\beta}))$, within a segment. In vector form, we have

$$\nabla_{\mathcal{A}} f(\boldsymbol{\beta}) - \operatorname{sgn}(\nabla_{\mathcal{A}} f(\boldsymbol{\beta}))s(t) = \mathbf{0}. \tag{3.1}$$

In general $s(t)$ can be any smooth and monotonic function that decreases from $s(0)$ to $s(t_{\max}) = 0$ at some finite $t_{\max} > 0$, where $s(0) = \max_j |\nabla_j f(\mathbf{0})|$. Intuitively $s(t)$ controls how the common absolute score of active predictors decays with respect to solution index t. Different choices of $s(t)$ lead to different indexing systems yet the same solution path. By construction $s(t)$ is larger than the absolute value of the scores of inactive predictors which are packed at zero. Once the absolute score $|\nabla_j f(\boldsymbol{\beta}(t))|$ of an inactive predictor coincides with $s(t)$, it joins the club and its score conforms to the ruling $s(t)$ thereafter. The classical LAR (Efron et al., 2004) sets $s(t) = s(0) - t$ which implies that $t_{\max} = s(0) = \max_j |\nabla_j f(\mathbf{0})|$. The following result follows from the implicit function theorem and provides the path following direction of the ConvexLAR algorithm within a path segment. See the Supplement for the proof and the following subsection for the definition of path segment operationally.

Theorem 3.1. *For a strictly convex and twice differentiable loss function $f(\boldsymbol{\beta})$, the LAR path solution $\boldsymbol{\beta}(t)$ is continuous and differentiable at t within a path segment. In addition, the solution vector $\boldsymbol{\beta}(t)$ satisfies the ordinary differential equation*

$$\frac{d}{dt}\boldsymbol{\beta}_{\mathcal{A}}(t) \quad = \quad s'(t)\boldsymbol{H}_{\mathcal{A}}^{-1}(\boldsymbol{\beta})\operatorname{sgn}(\nabla_{\mathcal{A}} f(\boldsymbol{\beta}(t))) \tag{3.2}$$

and $\beta_j(t) = 0$ for any $j \notin \mathcal{A}_t$.

3.2.1 ConvexLAR Algorithm

Initialize: $s(0) = \max_j |\nabla_j f(\mathbf{0})|$, $\boldsymbol{\beta}(0) = \mathbf{0}$, and $\mathcal{A} = \text{argmax}_j |\nabla_j f(\mathbf{0})|$

repeat

  Solve ODE

$$\frac{d}{dt}\boldsymbol{\beta}_{\mathcal{A}}(t) = s'(t)\boldsymbol{H}_{\mathcal{A}}^{-1}(\boldsymbol{\beta}(t))\text{sgn}(\nabla_{\mathcal{A}} f(\boldsymbol{\beta}(t))$$

  until $|\nabla_j f(\boldsymbol{\beta}(t))| = s(t)$ for some $j \notin \mathcal{A}$

  Update set $\mathcal{A} \leftarrow \mathcal{A} \cup \{j : |\nabla_j f(\boldsymbol{\beta}(t))| = s(t), j \notin \mathcal{A}\}$

until $s(t) = 0$ *or* $\boldsymbol{H}_{\mathcal{A}}(\boldsymbol{\beta}(t))$ *is singular*

Algorithm 1: ConvexLAR.

Theorem 3.1 suggests that the exact solution path of LAR can be obtained by solving the simple ODE system (3.2) segment by segment. The size of the ODE system within a segment is equal to the corresponding number of active predictors $|\mathcal{A}|$. The ConvexLAR algorithm is summarized in Algorithm 1. We initialize our solution path with $\boldsymbol{\beta}(0) = \mathbf{0}$ and the beginning active set contains the predictors that change the objective function $f(\mathbf{0})$ fastest. That is

$$\mathcal{A}_0 = \{\text{argmax}_j |\nabla_j f(\mathbf{0})|\}.$$

We then follow the solution path by solving the ODE system (3.2) until one or more new variables join the active set at some $t_1 > 0$, which is determined by the moment the active score $s(t)$ matches the maximum (absolute) gradient of one or more non-active predictors. The active predictor set $\mathcal{A}_t$ stays the same within $t \in [t_0, t_1)$. At t_1, it is updated by adding the new predictors that newly join the club. This process continues segment by segment until all the predictors are active. Then, in the final segment, the ConvexLAR solution path moves along a direction such that the absolute values of the first-order partial derivatives decrease at the same speed to zero, which happens at $t_{\max}$. Under assumptions of Theorem 3.1, the solution $\boldsymbol{\beta}(t_{\max})$ is the global minimizer of the convex loss function, just like the LAR solution ends at the full ordinary least squares estimate. This completes our ConvexLAR solution path algorithm.

Remark 3.1. *For the specific choice of* $s(t) = s(0) - t$*,*

$$t_1 = s(0) - s(t_1) = \left(\max_{j'} |\nabla_{j'} f(\mathbf{0})| \right) - |\nabla_j f(\boldsymbol{\beta}(t_1))|$$

for any $j \in \mathcal{A}_{t_1}$*. This holds analogously for later transition points. At end of the* m*-th LAR segment, the transition point*

$$t_m = \left(\max_{j'} |\nabla_{j'} f(\mathbf{0})| \right) - |\nabla_j f(\boldsymbol{\beta}(t_m))|$$

for any $j \in \mathcal{A}_{t_m}$*.*

Remark 3.2. *For least squares problem, the loss is a quadratic function* $f(\boldsymbol{\beta}) = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2/2$ *with a constant Hessian matrix* $\boldsymbol{H}(\boldsymbol{\beta}) = \boldsymbol{X}^\top \boldsymbol{X}$*. Since* $\mathrm{sgn}(\nabla_j f(\boldsymbol{\beta}(t)))$ *is constant for all* $j \in \mathcal{A}_t$ *in a neighborhood of* t*,* $\boldsymbol{H}_{\mathcal{A}}^{-1}(\boldsymbol{\beta}) \,\mathrm{sgn}(\nabla_{\mathcal{A}} f(\boldsymbol{\beta}(t)))$ *in* (3.2) *is piecewise constant. This leads to the piecewise linear solution path of the original LAR (Efron et al., 2004).*

Remark 3.3 (Non-strictly convex losses)**.** *The strict convexity assumption on the loss* f *precludes applications with non-convex losses or the* $n < p$ *least squares case. However from the proof in Supplement, we observe that the only essential ingredient is the positive definiteness of* $\boldsymbol{H}_{\mathcal{A}}(\boldsymbol{\beta}(t))$*. Therefore in non-strictly convex cases, we terminate path following as soon as* $\boldsymbol{H}_{\mathcal{A}}(\boldsymbol{\beta}(t))$ *becomes singular.*

Remark 3.4 (Partial regularization)**.** *So far we have assumed that the full set of parameters* $\boldsymbol{\beta}$ *are subject to regularization. In many applications, only a subset of parameters are regularized. Assume that the loss takes the form* $f(\boldsymbol{\beta}_0, \boldsymbol{\beta})$*, where* $\boldsymbol{\beta}_0 \in \mathbb{R}^{p_0}$ *is the vector of parameters exempt from regularization. Depending on the objective function, it may not be easy to remove* $\boldsymbol{\beta}_0$ *from the objective. However we can always define the marginal minimizer of* $\boldsymbol{\beta}_0$ *as a function of* $\boldsymbol{\beta}$

$$\boldsymbol{\beta}_0(\boldsymbol{\beta}) = \mathrm{argmin}_{\boldsymbol{\beta}_0} f(\boldsymbol{\beta}_0, \boldsymbol{\beta}). \tag{3.3}$$

Assume that f is strictly convex and thrice differentiable, then this mapping is uniquely defined and twice differentiable by the implicit function theorem. Denote the Jacobian and Hessian of this mapping by $D\boldsymbol{\beta}_0(\boldsymbol{\beta}) \in \mathbb{R}^{p_0 \times p}$ and $H\boldsymbol{\beta}_0(\boldsymbol{\beta}) \in \mathbb{R}^{p_0 p \times p}$ respectively. Then the first two derivatives with respect to $\boldsymbol{\beta}$ required in Theorem 3.1 can be obtained by the chain rule

$$\begin{aligned}\nabla f(\boldsymbol{\beta}) &= \nabla_{\boldsymbol{\beta}} f(\boldsymbol{\beta}_0, \boldsymbol{\beta}) + \nabla_{\boldsymbol{\beta}_0} f(\boldsymbol{\beta}_0, \boldsymbol{\beta}) \cdot D\boldsymbol{\beta}_0(\boldsymbol{\beta}), \\ Hf(\boldsymbol{\beta}) &= d^2_{\boldsymbol{\beta},\boldsymbol{\beta}} f(\boldsymbol{\beta}_0, \boldsymbol{\beta}) + D\boldsymbol{\beta}_0(\boldsymbol{\beta})^{\mathsf{T}} \cdot d^2_{\boldsymbol{\beta}_0,\boldsymbol{\beta}_0} f(\boldsymbol{\beta}_0, \boldsymbol{\beta}) \cdot D\boldsymbol{\beta}_0(\boldsymbol{\beta}) + [d_{\boldsymbol{\beta}_0} f(\boldsymbol{\beta}_0, \boldsymbol{\beta}) \otimes \boldsymbol{I}_p] H\boldsymbol{\beta}_0(\boldsymbol{\beta}).\end{aligned}$$

The Ada-Boost and Gaussian graphical model examples in Section 3.4 illustrate this strategy.

3.2.2 ConvexLASSO Modification

Efron et al. (2004) showed that in the least squares case the lasso solution path can be obtained by a slight modification of the LAR. Same extension applies to ConvexLAR. Consider the lasso regularized problem

$$\min_{\boldsymbol{\beta}} \quad f(\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} |\beta_j|. \tag{3.4}$$

With $\lambda = s(t) = s(0) - t$, the optimality condition for lasso solution is

$$\begin{aligned}\nabla_j f(\boldsymbol{\beta}) + \lambda \operatorname{sgn}(\beta_j) &= 0, \quad \beta_j \neq 0 \\ |\nabla_j f(\boldsymbol{\beta})| &\leq \lambda, \quad \beta_j = 0.\end{aligned} \tag{3.5}$$

A proof similar to that for Theorem 3.1 shows that the lasso solution path moves along the direction

$$\begin{aligned}\frac{d}{dt}\boldsymbol{\beta}_{\mathcal{A}}(t) &= -s'(t)\boldsymbol{H}_{\mathcal{A}}(\boldsymbol{\beta}(t)) \operatorname{sgn}(\boldsymbol{\beta}_{\mathcal{A}}(t)) \\ &= s'(t)\boldsymbol{H}_{\mathcal{A}}(\boldsymbol{\beta}(t)) \operatorname{sgn}(\nabla_{\mathcal{A}} f(\boldsymbol{\beta}(t)))\end{aligned}$$

until either (i) $\beta_j(t)$ hits zero for an active predictor j or (ii) $|\nabla_j f(\boldsymbol{\beta}(t))|$ hits boundary $\lambda = s(t)$

for some inactive predictor j. Both events change the active set and redefine the direction. The second equation is based on the fact that by (3.5) a lasso regularized solution satisfies $\mathrm{sgn}(\boldsymbol{\beta}_{\mathcal{A}}(t)) = -\mathrm{sgn}(\nabla_{\mathcal{A}} f(\boldsymbol{\beta}(t)))$. Similarly, the ConvexLAR algorithm can be modified to obtain the lasso solution path $\boldsymbol{\beta}(\lambda)$ with $\lambda = s(t)$. Observe that the event defining the LAR segment, i.e., the gradient of an inactive predictor hits $\lambda = s(t)$, is exactly same as the second type of event for lasso path. However the first type of event, i.e., the coefficient of an active predictor hitting zero, is not tracked in ConvexLAR. We call the modified algorithm, which tracks both types of events, by ConvexLASSO and the pseudocode is listed in Algorithm 2. Same argument as in Efron et al. (2004) and Wu (2011) shows that, under the assumption that, at each transition point, only one single event can happen, namely either one inactive predictor variable becomes active or one currently active predictor variable becomes inactive, ConvexLASSO algorithm yields the lasso solution path.

Initialize: $s(0) = \max_j |\nabla_j f(\mathbf{0})|$, $\boldsymbol{\beta}(0) = \mathbf{0}$, and $\mathcal{A} = \mathrm{argmax}_j |\nabla_j f(\mathbf{0})|$

repeat

  Solve ODE

$$\frac{d}{dt}\boldsymbol{\beta}_{\mathcal{A}}(t) = s'(t)\boldsymbol{H}_{\mathcal{A}}^{-1}(\boldsymbol{\beta}(t))\mathrm{sgn}(\nabla_{\mathcal{A}} f(\boldsymbol{\beta}(t))$$

  until (i) $\beta_j(t) = 0$ for some $j \in \mathcal{A}$ or (ii) $|\nabla_j f(\boldsymbol{\beta}(t))| = s(t)$ for some $j \notin \mathcal{A}$

  if *event (i)* **then**

    Update set $\mathcal{A} \leftarrow \mathcal{A} \setminus \{j : \beta_j(t) = 0, j \in \mathcal{A}\}$

  else if *event (ii)* **then**

    Update set $\mathcal{A} \leftarrow \mathcal{A} \cup \{j : |\nabla_j f(\boldsymbol{\beta}(t))| = s(t), j \notin \mathcal{A}\}$

  end

until $s(t) = 0$ *or* $\boldsymbol{H}_{\mathcal{A}}(\boldsymbol{\beta}(t))$ *is singular*

Algorithm 2: ConvexLASSO.

3.3 Generalizations

In this section we first summarize a few essential features of ConvexLAR and then demonstrate how these aspects lead to various generalizations.

L1. The "influence" of each predictor on the loss function is measured by the magnitude of its score (gradient). Indeed it is these gradient (score) functions that ConvexLAR operates on. For the ConvexLAR algorithm to be properly working, we require the influence function to be a monotone $\mathbb{R}^p \to \mathbb{R}^p$ mapping (Ortega and Rheinboldt, 2000). For instance, convexity of a loss function guarantees that its gradient (score) function is a monotone mapping.

L2. Certain form of "democratic voting" is enforced among the active predictors. Both the original LAR and ConvexLAR force the influences of individual active predictors to be equal. This equality constraint can be generalized when we want to favor certain predictors over others or to impose group structure on predictors.

L3. The influences of active predictors continuously decrease along the path so that those of inactive predictors can catch up. The assumption in **L1** that the gradient (score) function must be a monotone mapping guarantees that all influences change continuously.

L4. The inactive predictors keep parked at zero until their influence meets that of active ones, at which point they join the club.

L5. The influences of all active predictors gradually decrease at the same rate and hit zero at the same time, which declares the end of path following.

3.3.1 Weighted/Adaptive ConvexLAR

In many applications, there exists prior information about the importance of predictors, which can also be obtained in a data driven fashion. This motivates the development of adaptive lasso (Zou, 2006), which enjoys favorable asymptotic properties. To incorporate such information into ConvexLAR, we weight the "influence" of each predictor differentially and consider the weighted "influence" $w_j|\nabla_j f(\boldsymbol{\beta})|$ of each predictor, where $w_j \geq 0$ are predictor specific weights. A larger w_j implies higher "influence" and vice versa. A zero weight means no regularization for the corresponding predictor. For simplicity, we assume that all weights are positive. In this case the function $s(t)$ is the common value $w_j|\nabla_j f(\boldsymbol{\beta})|$ shared by active predictors. Setting $w_j \equiv 1$ reduces to the ConvexLAR. The stationarity condition for the "democratic voting" reads

$$\nabla_j f(\boldsymbol{\beta}) - w_j^{-1}\text{sgn}(\nabla_j f(\boldsymbol{\beta}))s(t) = 0, \quad j \in \mathcal{A},$$

and the ODE system becomes

$$\frac{d}{dt}\boldsymbol{\beta}_{\mathcal{A}}(t) = s'(t)\boldsymbol{H}_{\mathcal{A}}^{-1}(\boldsymbol{\beta})\text{diag}(\boldsymbol{w}_{\mathcal{A}}^{-1})\text{sgn}(\nabla_{\mathcal{A}} f(\boldsymbol{\beta}(t))),$$

where the vector $\boldsymbol{w}_{\mathcal{A}}^{-1}$ collects the inverse weights w_j^{-1} for active predictors. Segment terminates when $w_j|\nabla_j f(\boldsymbol{\beta}(t))|$ hits $s(t)$ for some inactive predictors $j \notin \mathcal{A}$. The weighted ConvexLAR is summarized in Algorithm 3 and a similar modification can be applied to get the corresponding adaptive ConvexLASSO.

1 Initialize $s(0) = \max_j (w_j|\nabla_j f(\mathbf{0})|)$, $\boldsymbol{\beta}(0) = \mathbf{0}$, and $\mathcal{A} = \text{argmax}_j w_j|\nabla_j f(\mathbf{0})|$
2 **repeat**
3 Solve ODE

$$\frac{d}{dt}\boldsymbol{\beta}_{\mathcal{A}}(t) = s'(t)\boldsymbol{H}_{\mathcal{A}}^{-1}(\boldsymbol{\beta}(t))\text{diag}(\boldsymbol{w}_{\mathcal{A}}^{-1})\text{sgn}(\nabla_{\mathcal{A}} f(\boldsymbol{\beta}(t))$$

until $w_j|\nabla_j f(\boldsymbol{\beta}(t))| = s(t)$ for some $j \notin \mathcal{A}$
4 Update set $\mathcal{A} \leftarrow \mathcal{A} \cup \{j : w_j|\nabla_j f(\boldsymbol{\beta}(t))| = s(t), j \notin \mathcal{A}\}$
5 **until** $s(t) = 0$ *or* $\boldsymbol{H}_{\mathcal{A}}(\boldsymbol{\beta}(t))$ *is singular*

Algorithm 3: Weighted ConvexLAR with predictor specific weights $w_j > 0$.

3.3.2 Group ConvexLAR

In this section we outline a strategy for extending ConvexLAR to incorporate group structure among predictors. Suppose predictors are divided into m groups with group size p_i, $i = 1, \ldots, m$. Slightly abusing notation, we use g to represent both the g-th group and the index set of all predictors belonging to the g-th group. In a similar manner, we use $\mathcal{G}$ to represent both the set of active groups and the index set of all predictors belonging to current active groups. For an arbitrary matrix $\boldsymbol{H}$, $\boldsymbol{H}_{\mathcal{I},\mathcal{J}}$ denotes the sub-matrix of $\boldsymbol{H}$ with rows in $\mathcal{I}$ and columns in $\mathcal{J}$.

Group ConvexLAR can be devised based on the following considerations. (i) We need one way to gauge the aggregate "influence", denoted by $I_g(\boldsymbol{\beta})$, of a group g. For instance, we can use either the ℓ_2 norm or ℓ_1 norm of the sub-vector of the gradient corresponding to a group, i.e., $I_g(\boldsymbol{\beta}) = \|\nabla_g f(\boldsymbol{\beta})\|_2 = \sqrt{\sum_{j\in g}[\nabla_j f(\boldsymbol{\beta})]^2}$ or $I_g(\boldsymbol{\beta}) = \|\nabla_g f(\boldsymbol{\beta})\|_1 = \sum_{j\in g}|\nabla_j f(\boldsymbol{\beta})|$. Note that in principle any ℓ_p norm, $p \geq 1$, can be used. (ii) For all groups in the active group set $\mathcal{G}$, we use the function $s(t)$ to control the common value of $e(p_g)^{-1}I_g(\boldsymbol{\beta})$, $g \in \mathcal{G}$, where p_g denotes the number of predictors in group g and $e(p_g)$ is the effective size of g-th group. Common choices for $e(p_g)$ are $\sqrt{p_g}$ or p_g. (iii) To make the ODE well defined, we need to assign the proportions of individual "influences" $c_g(j)$ to each predictor j in group g. We require $\|\boldsymbol{c}_g\|_p = 1$ to satisfy the equation $\|\nabla_g f(\boldsymbol{\beta})\|_p = \|\boldsymbol{c}_g I_g(\boldsymbol{\beta})\|_p$, where $\boldsymbol{c}_g$ is a column vector with $c_g(j)$, $j \in g$, stacked together. These conditions imply the general group LAR identity

$$\nabla_j f(\boldsymbol{\beta}) = c_g(j)I_g(\boldsymbol{\beta}) = c_g(j)e(p_g)s(t), \quad j \in g, g \in \mathcal{G}. \tag{3.6}$$

Now the implicit function theorem yields the group LAR updating direction (3.7) in Algorithm 4. Here $\boldsymbol{c}_{\mathcal{G}}$ is a column vector with $\boldsymbol{c}_g$ for $g \in \mathcal{G}$ stacked together and $\boldsymbol{p}_{\mathcal{G}}$ is a column vector with the corresponding effective group size $e(p_g)$ for each predictor j in group g stacked together.

Specific choices of the group "influence" $I_g(\boldsymbol{\beta})$, effective group size $e(p_g)$, and individual "influence" $c_g(j)$ lead to different updating directions. We specialize to the following three variants. The first has a close connection to the group lasso. The second and third recover two

Initialize $s(0) = \max_g e^{-1}(p_g) I_g(\mathbf{0})$, $\boldsymbol{\beta}(0) = \mathbf{0}$, and $\mathcal{G} = \operatorname{argmax}_g \frac{I_g(\mathbf{0})}{e(p_g)}$

repeat

Solve ODE system defined by

$$\frac{d}{dt}\boldsymbol{\beta}_{\mathcal{G}}(t) = \boldsymbol{H}_{\mathcal{G}}^{-1}(\boldsymbol{\beta}(t)) \left[s'(t)\mathrm{diag}(\boldsymbol{p}_{\mathcal{G}})\boldsymbol{c}_{\mathcal{G}} + s(t)\mathrm{diag}(\boldsymbol{p}_{\mathcal{G}})\frac{d}{dt}\boldsymbol{c}_{\mathcal{G}}(t) \right] \tag{3.7}$$

and $\boldsymbol{\beta}_g = \mathbf{0}$ for any inactive group $g \notin \mathcal{G}$ until $I_{g^*}(\boldsymbol{\beta}) = s(t)e(p_{g^*})$ for some $g^* \notin \mathcal{G}$

Update set $\mathcal{G} \leftarrow \mathcal{G} \cup \{g^*\}$

until $s(t) = 0$ *or* $\boldsymbol{H}_{\mathcal{G}}(\boldsymbol{\beta}(t))$ *is singular*

Algorithm 4: A general scheme for Group ConvexLAR.

versions of group LAR algorithms in literature.

1. **GroupConvexLAR:** We measure the group influence by $I_g(\boldsymbol{\beta}) = \|\nabla_g f(\boldsymbol{\beta})\|_2$, which is distributed to individual predictors within the group according to their effect size $c_g(j) = -\beta_j/\|\boldsymbol{\beta}_g\|_2$. Let $e(p_g) = \sqrt{p_g}$ and assume $\mathcal{G} = \mathcal{G}_0 \cup \mathcal{G}_1$, where $\mathcal{G}_0 = \{g_0, \ldots, g_a\}$ denotes the set of active groups with $\|\boldsymbol{\beta}_g\|_2 = 0$ and $\mathcal{G}_1 = \{g_{a+1}, \ldots, g_{a+b}\}$ denotes the set of active groups with $\|\boldsymbol{\beta}_g\|_2 > 0$.

Theorem 3.2. *For a strictly convex and twice differentiable loss function $f(\boldsymbol{\beta})$, the LAR solution $\boldsymbol{\beta}(t)$ is continuous and differentiable at t within a segment.*

(a) If $\mathcal{G}_0$ is an empty set, the active solution vector $\boldsymbol{\beta}_{\mathcal{G}}(t)$ satisfies the differential equation

$$\begin{aligned} \frac{d}{dt}\boldsymbol{\beta}_{\mathcal{G}}(t) &= [\boldsymbol{H}_{\mathcal{G}}(t) + \boldsymbol{D}]^{-1} s'(t)\mathrm{diag}(\boldsymbol{p}_{\mathcal{G}})\boldsymbol{c}_{\mathcal{G}} \\ &= \frac{s'(t)}{s(t)}[\boldsymbol{H}_{\mathcal{G}}(t) + \boldsymbol{D}]^{-1}\nabla_{\mathcal{G}} f(\boldsymbol{\beta}), \end{aligned} \tag{3.8}$$

where $\boldsymbol{D}$ is the block diagonal matrix with blocks

$$\frac{s(t)\sqrt{p_g}}{\|\boldsymbol{\beta}_g(t)\|_2}\left(\boldsymbol{I}_{p_g} - \frac{1}{\|\boldsymbol{\beta}_g(t)\|_2^2}\boldsymbol{\beta}_g(t)\boldsymbol{\beta}_g^{\intercal}(t)\right), \quad g \in \mathcal{G}.$$

(b) If $\mathcal{G}_0$ is not empty, the solution vector $\boldsymbol{\beta}_{g_i}$ for $g_i \in \mathcal{G}_0$ satisfies

$$\frac{d\boldsymbol{\beta}_{g_i}(t)}{dt} = k_i \nabla_{g_i} f(\boldsymbol{\beta}), \tag{3.9}$$

and the constants k_i, $i = 1, \ldots, a$, and updating direction for the groups in $\mathcal{G}_1$ are jointly determined by

$$\begin{pmatrix} k_1 \\ \vdots \\ k_a \\ \frac{d}{dt}\boldsymbol{\beta}_{\mathcal{G}_1}(t) \end{pmatrix} = s(t) \begin{pmatrix} \boldsymbol{A} & \boldsymbol{B}^\top \\ \boldsymbol{B} & \boldsymbol{C} \end{pmatrix}^{-1} \begin{pmatrix} p_{g_1} s'(t) \\ \vdots \\ p_{g_a} s'(t) \\ \nabla_{g_{a+1}} f(\boldsymbol{\beta})/s'(t) \\ \vdots \\ \nabla_{g_{a+b}} f(\boldsymbol{\beta})/s'(t) \end{pmatrix}, \tag{3.10}$$

where $\boldsymbol{A} \in \mathbb{R}^{a \times a}$ has entries $a_{ij} = d_{g_i} f(\boldsymbol{\beta}) H_{g_i,g_j}(\boldsymbol{\beta}) \nabla_{g_j} f(\boldsymbol{\beta})$, $1 \leq i, j \leq a$,

$$\boldsymbol{B} = \left[\boldsymbol{H}_{\mathcal{G}_1,g_1} \nabla_{g_1} f(\boldsymbol{\beta}), \quad \cdots, \quad \boldsymbol{H}_{\mathcal{G}_1,g_a} \nabla_{g_a} f(\boldsymbol{\beta}) \right] \in \mathbb{R}^{\sum_{i=a+1}^{a+b} p_i \times a},$$

and

$$\boldsymbol{C} = \boldsymbol{H}_{\mathcal{G}_1,\mathcal{G}_1} + \boldsymbol{D}_{\mathcal{G}_1,\mathcal{G}_1} \in \mathbb{R}^{\sum_{i=a+1}^{a+b} p_i \times \sum_{i=a+1}^{a+b} p_i}.$$

Note that, when $\mathcal{G}_0$ is empty, (3.10) reduces to (3.8). The technical proof of Theorem 3.2 is delegated to the Supplement. Again the strict convexity assumption can be relaxed to the positive definiteness of $\boldsymbol{H}_{\mathcal{G}}(\boldsymbol{\beta}(t))$ along the path, which guarantees the non-singularity of the matrix involved in (3.10).

2. **GroupConvexLAR-L1:** We choose $I_g(\boldsymbol{\beta}) = \|\nabla_g f(\boldsymbol{\beta})\|_1$ and $e(p_g) = p_g$. We further choose $c_g(j) = \frac{\nabla_j f(\boldsymbol{\beta}(t_g))}{\|\nabla_g f(\boldsymbol{\beta}(t_g))\|_1}$, where t_g is the time that the group g joins the active set $\mathcal{G}$. Note that in this case $c_g(j)$ is fixed for any $j \in g$ once group g joins the active set.

Therefore $\frac{d}{dt}\boldsymbol{c}_{\mathcal{G}}(t) = \mathbf{0}$ and the group LAR direction (3.7) reduces to

$$\frac{d}{dt}\boldsymbol{\beta}_{\mathcal{G}}(t) = s'(t)\boldsymbol{H}_{\mathcal{G}}(\boldsymbol{\beta})^{-1}\mathrm{diag}(\boldsymbol{p}_{\mathcal{G}})\boldsymbol{c}_{\mathcal{G}}. \tag{3.11}$$

3. **GroupConvexLAR-L2:** With the choice $I_g(\boldsymbol{\beta}) = \|\nabla_g f(\boldsymbol{\beta})\|_2$, $e(p_g) = \sqrt{p_g}$ and $c_g(j) = \frac{\nabla_j f(\boldsymbol{\beta}(t_g))}{\|\nabla_g f(\boldsymbol{\beta}(t_g))\|_2}$, the ODE updating direction is same as (3.11) with the obvious substitute for $\boldsymbol{c}_{\mathcal{G}}$ and $\boldsymbol{p}_{\mathcal{G}}$.

Note that, when all group sizes p_g are equal to 1,

$$\frac{\nabla_j f(\boldsymbol{\beta}(t_g))}{\|\nabla_g f\left(\boldsymbol{\beta}(t_g)\right)\|_2} = \frac{\nabla_j f(\boldsymbol{\beta}(t_g))}{\|\nabla_g f\left(\boldsymbol{\beta}(t_g)\right)\|_1} = \frac{-\beta_j}{\|\boldsymbol{\beta}_g\|_2} = \mathrm{sgn}(\nabla_j f(\boldsymbol{\beta}))$$

and $\frac{I_g(\boldsymbol{\beta})}{e(p_g)} = |\nabla_g f(\boldsymbol{\beta})|$ for all g. All three variants reduce to the ConvexLAR.

Connection with Previous Group LAR

Consider the variant GroupConvexLAR-L2 in the special case of least squares, i.e., $f(\boldsymbol{\beta}) = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2/2$. In this case, both $\boldsymbol{H}_{\mathcal{G}}(\boldsymbol{\beta}) = \boldsymbol{X}_{\mathcal{G}}^{\intercal}\boldsymbol{X}_{\mathcal{G}}$ and $\boldsymbol{c}_{\mathcal{G}}(t)$ are constant within a segment. Thus, the group LAR updating direction (3.11) is constant within each segment, leading to a piecewise linear solution path with segment-wise slope

$$\frac{d}{dt}\boldsymbol{\beta}_{\mathcal{G}}(t) = \frac{s'(t)}{s(t)}[\boldsymbol{X}_{\mathcal{G}}^{\intercal}\boldsymbol{X}_{\mathcal{G}}]^{-1}\nabla_{\mathcal{G}} f(\boldsymbol{\beta}).$$

This recovers a version of group LAR proposed by Yuan and Lin (2006) for group selection in least squares. Park and Hastie (2006) argue that this version of group LAR tends to select a large group with only few of its component correlated with the response. To avoid this problem, they propose another version of group LAR by simply replacing the average squared correlation $(\sum_{j\in g}[\nabla_j f(\boldsymbol{\beta})]^2/p_g)$ with the average absolute correlation $(\sum_{j\in g}|\nabla_j f(\boldsymbol{\beta})|/p_g)$, which is simply GroupConvexLAR-L1 specialized to least squares.

We emphasize that, for a general convex loss f, the solution paths by GroupConvexLAR-

L1 and GroupConvexLAR-L2 are both piecewise smooth instead of piecewise linear and ODE solving is necessary.

3.3.3 Group ConvexLASSO Modification

We show that a simple modification of first variant GroupConvexLAR yields the solution path of group lasso penalized problem

$$\min_{\boldsymbol{\beta}} \; f(\boldsymbol{\beta}) + \lambda \sum_{g} \sqrt{p_g} \|\boldsymbol{\beta}_g\|_2.$$

The Karush-Kuhn-Tucker (KKT) condition for the solution is

$$\nabla_g f(\boldsymbol{\beta}) + \lambda \frac{\sqrt{p_g}}{\|\boldsymbol{\beta}_g\|_2} \boldsymbol{\beta}_g = 0 \quad , \qquad \boldsymbol{\beta}_g \neq \mathbf{0}; \tag{3.12}$$

$$\|\nabla_g f(\boldsymbol{\beta})\|_2 \leq \lambda \sqrt{p_g} \quad , \qquad \boldsymbol{\beta}_g = \mathbf{0}. \tag{3.13}$$

With the choice $\lambda = s(t)$, the stationarity condition (3.12) coincides with the group LAR identity (3.6). This proves that the group lasso solution path moves along the direction (3.8) until either (i) all predictors of an active group hit zero or (ii) $\|\nabla_g f(\boldsymbol{\beta})\|_2$ hits boundary $\lambda\sqrt{p_g}$. Both events change the active set and redefine the direction. The second event is already considered in the GroupConvexLAR algorithm. Thus a simple modification of GroupConvexLAR leads to group lasso solution path, which is summarized in Algorithm 5. This exact path following algorithm for group lasso penalized convex loss seems new. There is no obvious connection between GroupConvexLAR-L1, L2 and group lasso.

3.4 Examples

We illustrate ConvexLAR and its extensions on various statistical problems. To demonstrate the efficiency of the proposed algorithm, we report the average running times of all numeric examples on i7 Core 2.93GHz, 8G RAM over 50 independent runs.

Initialize $s(0) = \max_g I_g(0)/\sqrt{p_g}$, $\boldsymbol{\beta}(0) = \mathbf{0}$, and $\mathcal{A} = \mathrm{argmax}_g I_g(0)/\sqrt{p_g}$
repeat
    Solve ODE according to (3.9) and (3.10) until
    (i) $\boldsymbol{\beta}_g(t) = 0$ for some $g \in \mathcal{A}$ or
    (ii) $I_g(\boldsymbol{\beta}) = s(t)\sqrt{p_g}$ for some $g \notin \mathcal{A}$
    if *event (i)* **then**
        Update set $\mathcal{A} \leftarrow \mathcal{A} \setminus \{g : \boldsymbol{\beta}_g(t) = 0, g \in \mathcal{A}\}$
    else if *event (ii)* **then**
        Update set $\mathcal{A} \leftarrow \mathcal{A} \cup \{g : I_g(\boldsymbol{\beta}) = s(t)\sqrt{p_g},\ g \notin \mathcal{A}\}$
    end
until $s(t) = 0$ *or* $\boldsymbol{H}_{\mathcal{G}}(\boldsymbol{\beta}(t))$ *is singular*

Algorithm 5: Group ConvexLASSO.

3.4.1 Recurrent event data

Suppose that we have n independent subjects in a recurrent event study. For each subject i, $N_i(t)$ denotes the number of events that occur over the interval $(0, t]$ and $\boldsymbol{x}_i \in \mathbb{R}^p$ is the corresponding covariate vector. Assume that given $\boldsymbol{x}_i$, the counting process $\{N_i(t)\}$ is a non-homogeneous Poisson process with mean function

$$\mu_i(t) = \mathrm{E}\{N_i(t)|\boldsymbol{x}_i\} = \mu_0(t)\exp(\boldsymbol{x}_i^\intercal\boldsymbol{\beta}) \tag{3.14}$$

for some unknown continuous baseline mean function $\mu_0(t)$ and unknown parameters $\boldsymbol{\beta} \in \mathbb{R}^p$. See Tong et al. (2009b) for a detailed introduction to recurrent event data.

As typical in survival study, each subject is subject to potential censoring. Let C_i denote the follow-up or dropout time for subject i and $\tilde{N}_i(t) = N_i(\min(t, C_i))$ be a point process for subject i's observed process. The observed data is summarized as $\{(\tilde{N}_i(t), C_i, \boldsymbol{x}_i), i = 1, 2, \cdots, n, 0 \leq t \leq T\}$, where the constant T denotes the maximum potential follow-up time. The log-partial likelihood function based on model (3.14) is given by

$$\ell(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\int_0^{\intercal}\left[\boldsymbol{x}_i^\intercal\boldsymbol{\beta} - \ln\left\{\sum_{j=1}^{n} Y_j(t)\exp(\boldsymbol{x}_j^\intercal\boldsymbol{\beta})\right\}\right] d\tilde{N}_i(t). \tag{3.15}$$

Under mild regularity conditions, the log-partial likelihood $\ell(\boldsymbol{\beta})$ is strictly concave in $\boldsymbol{\beta}$. Thus in this example, our objective function is chosen to be $f(\boldsymbol{\beta}) = -\ell(\boldsymbol{\beta})$. The first two derivatives of $f(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ are

$$
\begin{aligned}
\nabla f(\boldsymbol{\beta}) &= -\frac{1}{n}\sum_{i=1}^{n}\int_0^{\mathsf{T}}\left[\boldsymbol{x}_i - \frac{\sum_{j=1}^{n} Y_j(t)\exp(\boldsymbol{x}_j^{\mathsf{T}}\boldsymbol{\beta})\boldsymbol{x}_j}{\sum_{j=1}^{n} Y_j(t)\exp(\boldsymbol{x}_j^{\mathsf{T}}\boldsymbol{\beta})}\right] d\tilde{N}_i(t),\\
d^2 f(\boldsymbol{\beta}) &= \frac{1}{n}\sum_{i=1}^{n}\int_0^{\mathsf{T}}\left[\frac{\sum_{j=1}^{n} Y_j(t)\exp(\boldsymbol{x}_j^{\mathsf{T}}\boldsymbol{\beta})\boldsymbol{x}_j\boldsymbol{x}_j^{\mathsf{T}}}{\sum_{j=1}^{n} Y_j(t)\exp(\boldsymbol{x}_j^{\mathsf{T}}\boldsymbol{\beta})} - \right.\\
&\left.\frac{\left(\sum_{j=1}^{n} Y_j(t)\exp(\boldsymbol{x}_j^{\mathsf{T}}\boldsymbol{\beta})\boldsymbol{x}_j\right)\left(\sum_{j=1}^{n} Y_j(t)\exp(\boldsymbol{x}_j^{\mathsf{T}}\boldsymbol{\beta})\boldsymbol{x}_j\right)^{\mathsf{T}}}{\left(\sum_{j=1}^{n} Y_j(t)\exp(\boldsymbol{x}_j^{\mathsf{T}}\boldsymbol{\beta})\right)^2}\right] d\tilde{N}_i(t).
\end{aligned}
$$

Tong et al. (2009b) studied a Chronic Granulomatous Disease (CGD) data which were collected from a multicenter placebo-controlled randomized trial of gamma inferon with chronic granulotomous disease. There were 128 patients randomized to two groups, gamma interferon group ($n_1 = 63$) and placebo group ($n_2 = 65$). For each patient the times from the beginning of the study to initial and any recurrent serious infections are available. Eleven covariates are considered: 1. trtmt=treatment (Yes/No), 2. inherit=pattern of inheritance (autosomal/recessive), 3. age, 4. height, 5. weight, 6. cortico=use of corticosteroids (Yes/No), 7. prophy=use of prophylactic antibiotics (Yes/No), 8. gender=female, 9. hosp1=hosp. (category: US/other), 10. hosp2=hosp. (category: Europe-Amsterdam), and 11. hosp3=hosp. (category: Europe-other). We standardize all continuous covariates (age, height and weight) to have mean 0 and unit length. Furthermore, we also include the six quadratic and interaction terms between three continuous covariates. They are: 12. age×age, 13. height×height, 14. weight×weight, 15. age×height, 16. age×weight, 17. height×weight.

Figure 3.1 shows the solution paths from different algorithms. Here the numbers on the right hand side indicate which variable each path corresponds to. In all plots, the x-axis are in the units of $\|\boldsymbol{\beta}(t)\|_1/\max_t\|\boldsymbol{\beta}(t)\|_1$ and vertical lines mark the event times for easy comparison between various solution paths. Top row of Figure 3.1 displays the ConvexLAR (top left panel) and ConvexLASSO (top right panel) solution paths for the CGD data. They are qualitatively different. For instance, in the ConvexLASSO path, β_9 hits zero and then escapes active set at Step 14. In addition we also apply the weighted CovexLAR and ConvexLASSO algorithms with weights set at the maximum likelihood estimator of equation (3.15). They differ significantly from the unweighted ones. The running times are 4.84, 5.60, 5.34 and 6.37 seconds for the ConvexLAR, ConvexLASSO, weighted ConvexLAR and weighted ConvexLASSO solution paths, respectively.

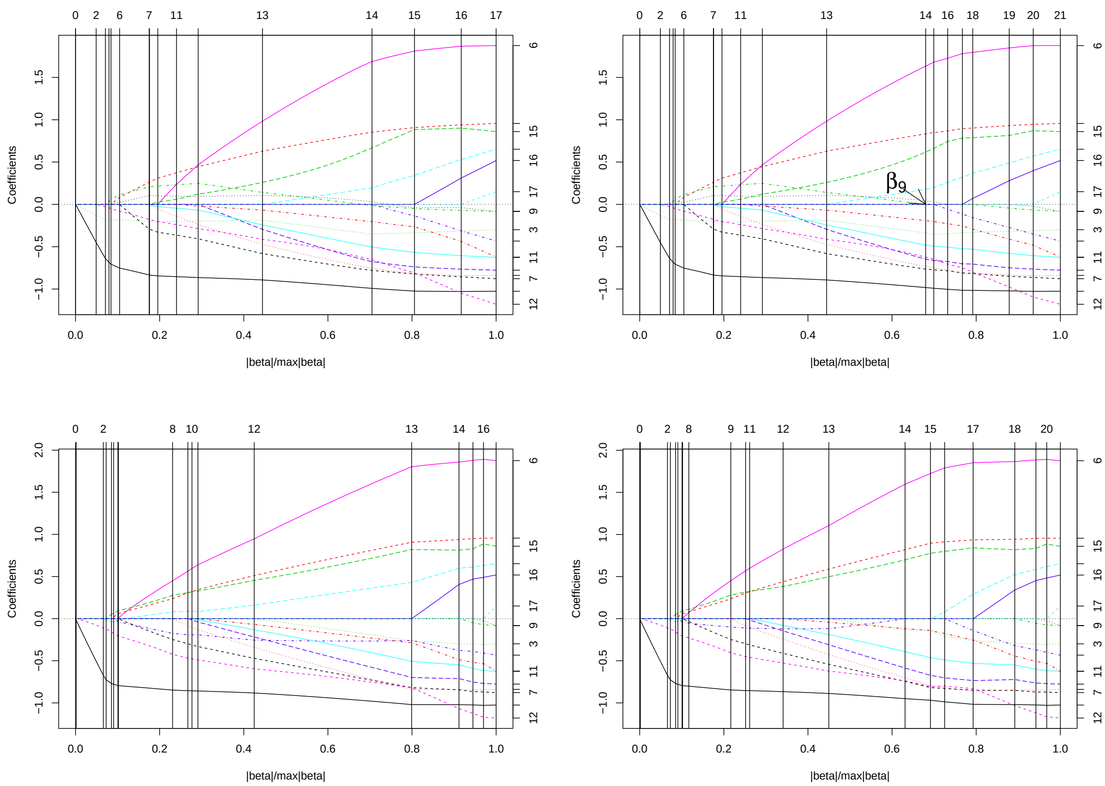


Figure 3.1: Recurrent event example (CGD data). Top left: ConvexLAR; Top right: ConvexLASSO; Bottom left: weighted ConvexLAR; Bottom right: weighted ConvexLASSO.

3.4.2 Panel Count Data

In the above recurrent event example, we assume that the exact time of each event, if not censored, is observed. Unfortunately this is not the case in many studies. The model for panel count data (Sun and Wei, 2000; Tong et al., 2009a) provides a remedy.

Let $T_{i1} < T_{i2} < \cdots < T_{im_i}$ be the potential observation times on process $N_i(t)$ and $H_i(t) = \sum_{j=1}^{m_i} I(T_{ij} \leq t)$ denote the observation process. Let $\tilde{H}_i(t) = H(\min(t, C_i))$ be the actual observation times after censoring. Then the observed data for panel count model is

$$\{(N_i(t)d\tilde{H}_i(t), \tilde{H}_i(t), C_i, \boldsymbol{x}_i^\mathsf{T}), \quad i = 1, \ldots, n, 0 \leq t \leq T\}.$$

When H_i and C_i are mutually independent and also independent of N_i and $\boldsymbol{x}_i$, Sun and Wei (2000) propose to estimate regression parameters by solving the following estimation equation

$$W(\boldsymbol{\beta}) = \sum_{i=1}^{n} \bar{N}_i e^{-\boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta}} \boldsymbol{x}_i = \mathbf{0},$$

where $\bar{N}_i = \int_0^\mathsf{T} N_i(t)d\tilde{H}_i(t)$. ConvexLAR and its extensions can be applied directly to the influence function

$$-W(\boldsymbol{\beta}) = \sum_{i=1}^{n} \bar{N}_i e^{-\boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta}} \boldsymbol{x}_i$$

which has a positive definite derivative

$$-DW(\boldsymbol{\beta}) = \sum_{i=1}^{n} \bar{N}_i e^{-\boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta}} \boldsymbol{x}_i \boldsymbol{x}_i^\mathsf{T}.$$

We illustrate with the bladder tumor recurrence data considered in Sun and Wei (2000). A total of 85 bladder tumor patients were randomized into two treatment groups, i.e., placebo group and thiotepa treatment group. Most patients visited the hospital several times to have their recurrent tumors removed. The number of new tumors discovered at each visit was recorded and removed after each visit. The binary treatment (trtmt) is one of our explanatory covariates. Furthermore, we consider two additional important baseline covariates, the number of initial tumors (num) and the size of the largest initial tumor (size). All covariates are centered around zero and scaled to have unit length. Figure 3.2 shows the ConvexLAR and ConvexLASSO solution paths for this example. The two solution paths are identical in this particular example as no active predictors return to zero along the path. The running times are 0.04 and 0.05 seconds for the ConvexLAR and ConvexLASSO solution paths, respectively.

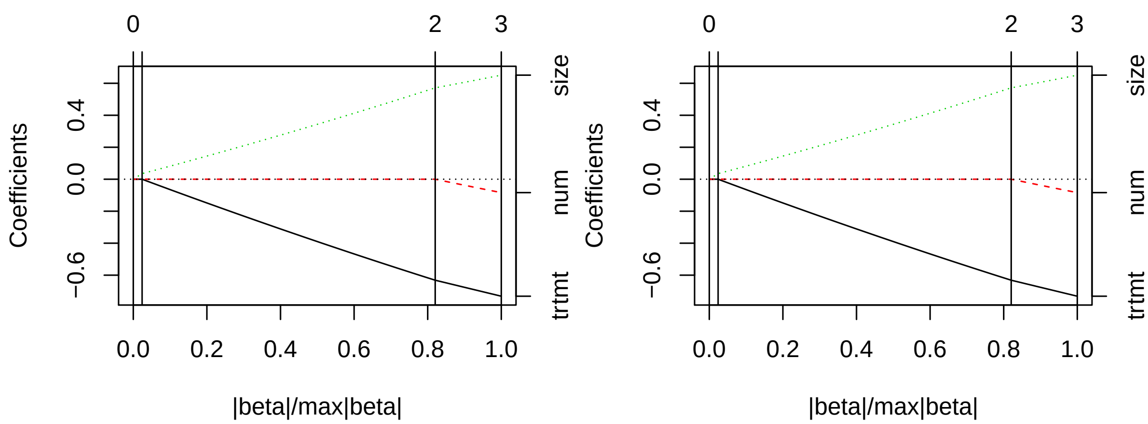


Figure 3.2: ConvexLAR and ConvexLASSO solution paths for the panel count data (bladder) example.

3.4.3 Ada-Boost

Ada-Boost is considered one of the best off-the-shelf classification methods (Hastie et al., 2009). In binary classification, we are given a training data set $\{(\boldsymbol{x}_i, y_i\} : i = 1, \cdots, n\}$ with $\boldsymbol{x}_i \in \mathbb{R}^p$ and $y_i \in \{-1, +1\}$. The goal is to estimate a function $f(\boldsymbol{x})$ whose sign will be used as the classification rule. For simplicity we consider the linear classification, in which $f(\boldsymbol{x}) = \beta_0 + \boldsymbol{x}^\mathsf{T}\boldsymbol{\beta}$.

The Ada-Boost estimates parameters by solving

$$\min_{\beta_0,\boldsymbol{\beta}} \sum_{i=1}^{n} e^{-y_i(\beta_0+\boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta})}, \tag{3.16}$$

which is strictly convex and thus amenable to our ConvexLAR algorithm. Denote $\mathcal{D} = \{i : y_i = -1\}$. Define the marginal minimizer of β_0 as a function of $\boldsymbol{\beta}$

$$\boldsymbol{\beta}_0(\boldsymbol{\beta}) = \text{argmin}_{\boldsymbol{\beta}_0} f(\boldsymbol{\beta}_0, \boldsymbol{\beta}) = \frac{1}{2} \log \left\{ \frac{\sum_{i \in \mathcal{D}^c} e^{-\boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta}}}{\sum_{i \in \mathcal{D}} e^{\boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta}}} \right\}. \tag{3.17}$$

Thus the first two derivatives of $f(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ are

$$\begin{aligned}
\nabla f(\boldsymbol{\beta}) &= \left(\frac{\partial \beta_0}{\partial \boldsymbol{\beta}}\right) \left[e^{\beta_0} \sum_{\mathcal{D}} e^{\boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta}} - e^{-\beta_0} \sum_{\mathcal{D}^c} e^{-\boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta}} \right] + \left[e^{\beta_0} \sum_{\mathcal{D}} e^{\boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta}} \boldsymbol{x}_i - e^{-\beta_0} \sum_{\mathcal{D}^c} e^{-\boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta}} \boldsymbol{x}_i \right], \\
d^2 f(\boldsymbol{\beta}) &= \left(\frac{\partial^2 \beta_0}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\mathsf{T}}\right) \left[e^{\beta_0} \sum_{\mathcal{D}} e^{\boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta}} - e^{-\beta_0} \sum_{\mathcal{D}^c} e^{-\boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta}} \right] \\
&\quad + \left(\frac{\partial \beta_0}{\partial \boldsymbol{\beta}}\right) \left(\frac{\partial \beta_0}{\partial \boldsymbol{\beta}}\right)^\mathsf{T} \left[e^{\beta_0} \sum_{\mathcal{D}} e^{\boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta}} + e^{-\beta_0} \sum_{\mathcal{D}^c} e^{-\boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta}} \right] \\
&\quad + \left(\frac{\partial \beta_0}{\partial \boldsymbol{\beta}}\right) \left[e^{\beta_0} \sum_{\mathcal{D}} e^{\boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta}} \boldsymbol{x}_i + e^{-\beta_0} \sum_{\mathcal{D}^c} e^{-\boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta}} \boldsymbol{x}_i \right]^\mathsf{T} \\
&\quad + \left[e^{\beta_0} \sum_{\mathcal{D}} e^{\boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta}} \boldsymbol{x}_i + e^{-\beta_0} \sum_{\mathcal{D}^c} e^{-\boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta}} \boldsymbol{x}_i \right] \left(\frac{\partial \beta_0}{\partial \boldsymbol{\beta}}\right)^\mathsf{T} \\
&\quad + \left[e^{\beta_0} \sum_{\mathcal{D}} e^{\boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta}} \boldsymbol{x}_i \boldsymbol{x}_i^\mathsf{T} + e^{-\beta_0} \sum_{\mathcal{D}^c} e^{-\boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta}} \boldsymbol{x}_i \boldsymbol{x}_i^\mathsf{T} \right],
\end{aligned}$$

where

$$\frac{\partial \beta_0(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -\frac{1}{2}\left\{\frac{\sum_{\mathcal{D}^c} e^{-\boldsymbol{x}_i^\top \boldsymbol{\beta}} \boldsymbol{x}_i}{\sum_{\mathcal{D}^c} e^{-\boldsymbol{x}_i^\top \boldsymbol{\beta}}} + \frac{\sum_{\mathcal{D}} e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}} \boldsymbol{x}_i}{\sum_{\mathcal{D}} e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}}}\right\},$$

$$\begin{aligned}\frac{\partial^2 \beta_0(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} =& \frac{1}{2}\left\{\frac{\left(\sum_{\mathcal{D}^c} e^{-\boldsymbol{x}_i^\top \boldsymbol{\beta}} \boldsymbol{x}_i \boldsymbol{x}_i^\top\right)\left(\sum_{\mathcal{D}^c} e^{-\boldsymbol{x}_i^\top \boldsymbol{\beta}}\right) - \left(\sum_{\mathcal{D}^c} e^{-\boldsymbol{x}_i^\top \boldsymbol{\beta}} \boldsymbol{x}_i\right)\left(\sum_{\mathcal{D}^c} e^{-\boldsymbol{x}_i^\top \boldsymbol{\beta}} \boldsymbol{x}_i\right)^\top}{\left(\sum_{\mathcal{D}^c} e^{-\boldsymbol{x}_i^\top \boldsymbol{\beta}}\right)^2}\right\} \\ & -\frac{1}{2}\left\{\frac{\left(\sum_{\mathcal{D}} e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}} \boldsymbol{x}_i \boldsymbol{x}_i^\top\right)\left(\sum_{\mathcal{D}} e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}}\right) - \left(\sum_{\mathcal{D}} e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}} \boldsymbol{x}_i\right)\left(\sum_{\mathcal{D}} e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}} \boldsymbol{x}_i\right)^\top}{\left(\sum_{\mathcal{D}} e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}}\right)^2}\right\}.\end{aligned}$$

The Wisconsin Diagnostic Breast Cancer (WDBC) data (Frank and Asuncion, 2010) are collected on $n = 569$ patients from digitized images of a fine needle aspirate (FNA) of their breast mass. The number of predictors is $p = 10$. The mean, standard error, and "worst" or largest (mean of the three largest values) of these predictors were computed for each patient, resulting in 30 features forming 10 groups each of size 3. The response is binary in that each patient is diagnosed either as malignant ($Y = 1$) or benign ($Y = -1$). Each predictor variable is standardized to have mean zero and variance one. Figure 3.3 displays the group ConvexLARs solution paths for this example. The x-axis is $\log(1 + s(t))$, where $s(t)$ is the same as the λ in group ConvexLASSO. To have a clear view, only the solution paths where $\log(1 + s(t)) > 3$ are plotted. GroupConvexLAR-L2 and GroupConvexLAR-L1 appear quite different from Group-ConvexLAR with larger $\max_{i\in\mathcal{G}} |\beta_i(t)|$ across the same level of $s(t)$. Bottom row of Figure 3.3 displays the GroupConvexLAR and Group ConvexLASSO solution paths, which are fundamentally different for the WDBC data. The third group (displayed with dotted line) is the first active one and then stay active along the whole GroupConvexLAR solution path. In contrast, the same group hits zero in the event 3 of the Group ConvexLASSO solution path and escapes the active set thereafter. The running times are 1.53, 1.58, 1.49 and 2.13 seconds for the GroupConvexLAR-L2, GroupConvexLAR-L1, GroupConvexLAR and GroupConvexLASSO solution paths, respectively.

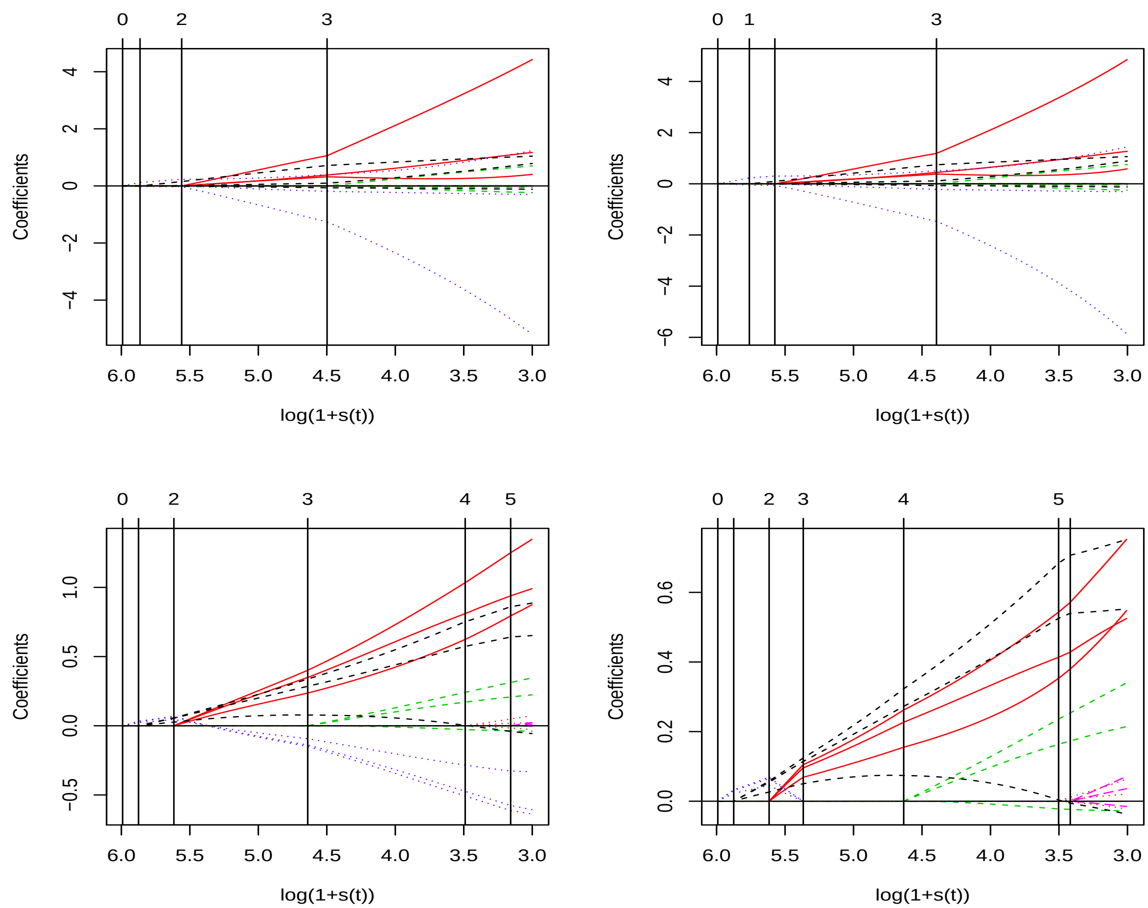


Figure 3.3: Group LARs solution paths for the Ada-Boost data (WDBC) example. Top left: GroupConvexLAR-L2; Top right: GroupConvexLAR-L1; Bottom left: GroupConvexLAR; Bottom right: Group ConvexLASSO.

3.4.4 Gaussian graphical model

Our fourth example concerns the LAR for Gaussian graphical model. Assume we have i-id observations $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n \in \mathbb{R}^p$ from $N(\mathbf{0}, \boldsymbol{\Sigma})$. Denote the sample variance covariance matrix by $\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i \boldsymbol{x}_i^\intercal$ and precision matrix by $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$. Then the negative likelihood for $\boldsymbol{\Omega}$ is given by $-\ell(\boldsymbol{\Omega}) = -\log\det\boldsymbol{\Omega} + \mathrm{tr}(\hat{\boldsymbol{\Sigma}}\boldsymbol{\Omega})$, which is convex. Let ω_{ij} denote the (i,j)-element of $\boldsymbol{\Omega}$. $w_{ij} = 0$, $i \neq j$, implies the conditional independence between variables i and j. We partition the parameters into the sets $\boldsymbol{\Omega}_0 = (\omega_{11}, \omega_{22}, \cdots, \omega_{pp})^\intercal \in \mathbb{R}^p$

and $\mathbf{\Omega}_1 = (\omega_{12}, \omega_{13}, \cdots, \omega_{(p-1)p})^\mathsf{T} \in \mathbb{R}^{p(p-1)/2}$. Only those in $\mathbf{\Omega}_1$ are subject to regularization. We may rewrite the negative loglikelihood as $-\ell(\mathbf{\Omega}_0, \mathbf{\Omega}_1)$. For every fixed $\mathbf{\Omega}_1$, we define $\mathbf{\Omega}_0(\mathbf{\Omega}_1) = \text{argmin}_{\mathbf{\Omega}_0} - \ell(\mathbf{\Omega}_0, \mathbf{\Omega}_1)$. With these notations, we have $f(\mathbf{\Omega}_1) = -\ell(\mathbf{\Omega}_0(\mathbf{\Omega}_1), \mathbf{\Omega}_1)$.

To derive the LAR solution path, we need the first two derivatives of the objective function

$$f(\mathbf{\Omega}_0, \mathbf{\Omega}_1) = -\log\det \mathbf{\Omega}(\mathbf{\Omega}_0, \mathbf{\Omega}_1) + \text{tr}(\hat{\mathbf{\Sigma}}\mathbf{\Omega}(\mathbf{\Omega}_0, \mathbf{\Omega}_1)),$$

where $\mathbf{\Omega}_0$ is implicitly determined by $\mathbf{\Omega}_1$. We first show how to determine $\mathbf{\Omega}_0$ given $\mathbf{\Omega}_1$. Setting partial derivative of f with respect to $\mathbf{\Omega}_0$

$$D_{\mathbf{\Omega}_0} f(\mathbf{\Omega}_0, \mathbf{\Omega}_1) = D_{\mathbf{\Omega}_0} f(\mathbf{\Omega}) D\mathbf{\Omega}(\mathbf{\Omega}_0) = -(\text{vec}\, \mathbf{\Omega}^{-1})^\mathsf{T} D\mathbf{\Omega}(\mathbf{\Omega}_0) + (\text{vec}\, \hat{\mathbf{\Sigma}})^\mathsf{T} D\mathbf{\Omega}(\mathbf{\Omega}_0)$$

to $\mathbf{0}$ gives the stationarity condition

$$\text{diag}(\mathbf{\Omega}^{-1}(\mathbf{\Omega}_0, \mathbf{\Omega}_1)) = (\hat{\sigma}_{11}, \ldots, \hat{\sigma}_{pp})^\mathsf{T}.$$

In other words, given $\mathbf{\Omega}_1$, we need to choose $\mathbf{\Omega}_0$ such that the diagonal entries of $\mathbf{\Omega}^{-1}$ match those of $\hat{\mathbf{\Sigma}}$. In practice, Newton's iteration

$$\omega_0^{(t+1)} = \omega_0^{(t)} + \left\{[D\mathbf{\Omega}(\mathbf{\Omega}_0)]^\mathsf{T}[\mathbf{\Omega}^{-1} \otimes \mathbf{\Omega}^{-1}][D\mathbf{\Omega}(\mathbf{\Omega}_0)]\right\}^{-1} [\text{diag}(\mathbf{\Omega}^{-1} - \hat{\mathbf{\Sigma}})]$$

can be applied to solve for $\mathbf{\Omega}_0$ given $\mathbf{\Omega}_1$. We denote this mapping by $\mathbf{\Omega}_0(\mathbf{\Omega}_1)$. The gradient $D\mathbf{\Omega}_0(\mathbf{\Omega}_1) \in \mathbb{R}^{p \times p(p-1)/2}$ will be of use later and is obtained through the implicit function theorem

$$D\mathbf{\Omega}_0(\mathbf{\Omega}_1) = -\left\{[D\mathbf{\Omega}(\mathbf{\Omega}_0)]^\mathsf{T}(\mathbf{\Omega}^{-1} \otimes \mathbf{\Omega}^{-1})[D\Omega(\mathbf{\Omega}_0)]\right\}^{-1} [D\mathbf{\Omega}(\mathbf{\Omega}_0)]^\mathsf{T}(\mathbf{\Omega}^{-1} \otimes \mathbf{\Omega}^{-1})[D\Omega(\mathbf{\Omega}_1)] \quad (3.18)$$

Now the first derivative of objective function f with respect to $\mathbf{\Omega}_1$ is

$$D_{\mathbf{\Omega}_1} f(\mathbf{\Omega}_0(\mathbf{\Omega}_1), \mathbf{\Omega}_1) = D_{\mathbf{\Omega}_1} f(\mathbf{\Omega}_0, \mathbf{\Omega}_1) + D_{\mathbf{\Omega}_0} f(\mathbf{\Omega}_0, \mathbf{\Omega}_1) D\mathbf{\Omega}_0(\mathbf{\Omega}_1)$$

but the second term vanishes because $D_{\mathbf{\Omega}_0} f(\mathbf{\Omega}_0, \mathbf{\Omega}_1) = \mathbf{0}$. Hence

$$D_{\mathbf{\Omega}_1} f(\mathbf{\Omega}_0(\mathbf{\Omega}_1), \mathbf{\Omega}_1) = Df(\mathbf{\Omega}) D\mathbf{\Omega}(\mathbf{\Omega}_1) = (-\operatorname{vec} \mathbf{\Omega}^{-1} + \operatorname{vec} \hat{\mathbf{\Sigma}})^{\mathsf{T}} D\mathbf{\Omega}(\mathbf{\Omega}_1).$$

In words, given current $\mathbf{\Omega}_1$, calculate $\mathbf{\Omega}^{-1}$ at optimal $\mathbf{\Omega}_0$; then the off-diagonal entries of $2(\hat{\mathbf{\Sigma}} - \mathbf{\Omega}^{-1})$ form the gradient of f in terms of $\mathbf{\Omega}_1$. For the Hessian,

$$\begin{aligned} Hf(\mathbf{\Omega}_1) &= D_{\mathbf{\Omega}_1}[D_{\mathbf{\Omega}_1} f(\mathbf{\Omega}_0(\mathbf{\Omega}_1), \mathbf{\Omega}_1)]^{\mathsf{T}} \\ &= D_{\mathbf{\Omega}_1}[D\mathbf{\Omega}(\mathbf{\Omega}_1)^{\mathsf{T}}(-\operatorname{vec} \mathbf{\Omega}^{-1} + \operatorname{vec} \hat{\mathbf{\Sigma}})] \\ &= -D_{\mathbf{\Omega}_1}[D\mathbf{\Omega}(\omega_1)^{\mathsf{T}}(\operatorname{vec} \mathbf{\Omega}^{-1})] \\ &= -[D\mathbf{\Omega}(\omega_1)]^{\mathsf{T}}(\mathbf{\Omega}^{-1} \otimes \mathbf{\Omega}^{-1})[D\mathbf{\Omega}(\mathbf{\Omega}_1) + D\mathbf{\Omega}(\mathbf{\Omega}_0) D\mathbf{\Omega}_0(\mathbf{\Omega}_1)]. \end{aligned} \tag{3.19}$$

Now substitute $D\mathbf{\Omega}_0(\mathbf{\Omega}_1)$ by the expression (3.18).

We illustrate our algorithm by two examples. The first data set contains 88 students' scores on five math courses - mechanics, vector, algebra, analysis and statistics. See (Mardia et al., 1979, Table 1.2.1) for more details. Figure 3.4 displays the ConvexLAR and ConvexLASSO solution paths. The most important three edges are analysis-algebra, statistics-algebra, and algebra-vector by lasso regularization. The ConvexLAR and ConvexLASSO solution paths coincide in this example. The running times are 0.28 and 0.37 seconds for the ConvexLAR and ConvexLASSO solution paths, respectively.

Our second example concerns a simulated data set with $p = 10$ and $n = 200$ and illustrates a case where ConvexLAR and ConvexLASSO yield different paths. The true precision matrix $\mathbf{\Omega} = (\omega_{ij})$ has entries $\omega_{ii} = 1$, $\omega_{i,i-1} = 0.5$, $\omega_{i,i+1} = 0.5$, and $\omega_{ij} = 0$ for $|i-j| > 1$. There are 45 non-diagonal free parameters. Figure 3.5 displays the ConvexLAR and ConvexLASSO solution paths. The solution paths appear different. The running times are 1.63 and 2.62 seconds for the ConvexLAR and ConvexLASSO solution paths, respectively.

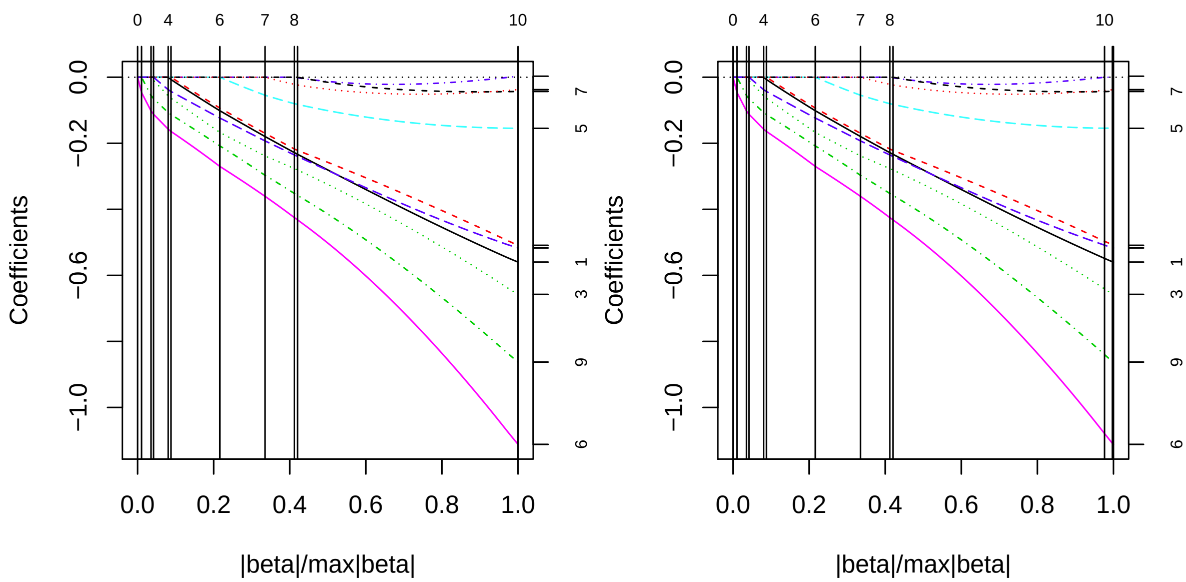


Figure 3.4: ConvexLAR and ConvexLASSO solution paths for the graphical model (math score) example.

3.5 Discussion

Variable selection has become an essential tool for modern data analysis. So far penalization method such as lasso has been the dominant regularization technique and extended to handle increasingly more applications. In contrast, the original LAR (Efron et al., 2004) has received much less generalizations despite its popularity. In this chapter, we show that the simple geometric idea in LAR can be naturally extended to various situations such as convex loss, group structures in predictors, and data adaptive regularizations. The classical score function plays an essential role throughout the development. The original "least angle" idea translates to the

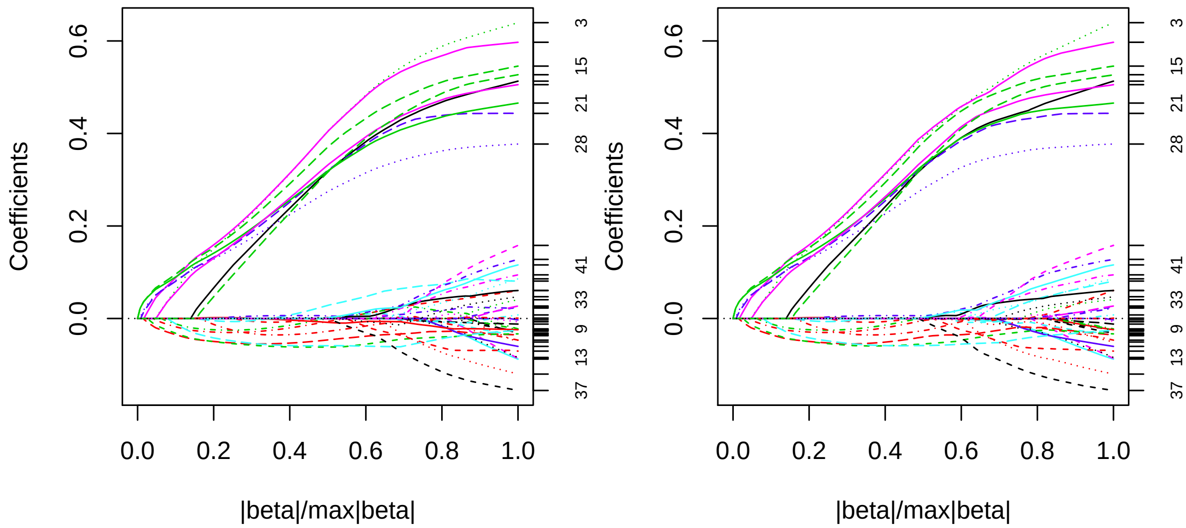


Figure 3.5: ConvexLAR and ConvexLASSO solution paths for the graphical model example with the simulated data.

equality of contributions by the active predictors to the score function. In our understanding, this is the fundamental idea in LAR and underpins the various extensions presented in this article.

This chapter is meant to whet readers' appetites not satiate them. Much is left undone. For instance, in principle it is the estimation equation that LAR operates on. Therefore LAR naturally applies many statistical methods without a natural loss function such as generalized estimation equation (GEE), as hinted in the panel count data in Section 3.4. In this chapter we focus on LAR algorithmic development and forego the theoretical treatment. There has been intensive study of the asymptotic properties of regularized estimates by penalty methods in recent years. Same study for ConvexLAR is worth pursuing. Especially those that might shed light on the difference between the two regularization methods are much desired.

3.6 Supplement

3.6.1 Proof of Theorem 3.1

The LAR fundamental identity (3.1) for active predictors dictates the vector equation

$$k(\boldsymbol{\beta}_{\mathcal{A}}, t) = \nabla_{\mathcal{A}} f(\boldsymbol{\beta}) - \operatorname{sgn}(\nabla_{\mathcal{A}} f(\boldsymbol{\beta}))s(t) = \mathbf{0}.$$

To solve for $\boldsymbol{\beta}_{\mathcal{A}}$ in terms of t, we apply the implicit function theorem (Lange, 2004). This requires calculating the differential of k with respect to the dependent variables $\boldsymbol{\beta}_{\mathcal{A}}$ and the independent variable t

$$\begin{aligned} \partial_{\boldsymbol{\beta}_{\mathcal{A}}} k(\boldsymbol{\beta}_{\mathcal{A}}, t) &= \boldsymbol{H}_{\mathcal{A}}(\boldsymbol{\beta}, t) \\ \partial_t k(\boldsymbol{\beta}_{\mathcal{A}}, t) &= -s'(t)\operatorname{sgn}(\nabla_{\mathcal{A}} f(\boldsymbol{\beta})). \end{aligned}$$

Given the non-singularity of $\boldsymbol{H}_{\mathcal{A}}(\boldsymbol{\beta}, t)$, the implicit function theorem applies and shows the continuity and differentiability of $\boldsymbol{\beta}_{\mathcal{A}}(t)$ at t. Furthermore, it supplies the derivative (3.2).

3.6.2 Derivation of GroupConvexLAR directions (3.8), (3.9) and (3.10)

We use $\mathcal{G}_0$ to denote the set of active groups that equal to $\mathbf{0}$. Let $\mathcal{G}_1 = \mathcal{G} - \mathcal{G}_0$, where $\mathcal{G}$ denotes the set of all active groups. We slightly abuse notation by letting $\mathcal{G}_0$, $\mathcal{G}_1$ and $\mathcal{G}$ to denote both the sets of groups and the set of all predictors belonging to the corresponding groups. Obviously, $\mathcal{G}_0$, $\mathcal{G}_1$ and $\mathcal{G}$ depend on the time index t.

Derivation of updating direction hinges upon the LAR identity (3.6), which we rewrite here for convenience

$$\nabla_g f(\boldsymbol{\beta}) + \sqrt{p_g}\, s(t) \frac{\boldsymbol{\beta}_g}{\|\boldsymbol{\beta}_g\|_2} = \mathbf{0}_{p_g}, \quad g \in \mathcal{G}. \tag{3.20}$$

1. When $\mathcal{G}_0$ is empty, differentiating the vector equation (3.20) with respect to t via chain

rule gives

$$\left[\boldsymbol{H}_{g,\mathcal{G}}(\boldsymbol{\beta})+\sqrt{p_g}\frac{s(t)}{\|\boldsymbol{\beta}_g\|_2}\left(\boldsymbol{I}_{p_g}-\frac{\boldsymbol{\beta}_g\boldsymbol{\beta}_g^{\mathsf{T}}}{\|\boldsymbol{\beta}_g\|_2^2}\right)D_{\boldsymbol{\beta}_{\mathcal{G}}}\boldsymbol{\beta}_g\right]\frac{d\boldsymbol{\beta}_{\mathcal{G}}(t)}{dt}+\sqrt{p_g}s'(t)\frac{\boldsymbol{\beta}_g}{\|\boldsymbol{\beta}_g\|}=\mathbf{0}_{p_g}, \quad (3.21)$$

where $D_{\boldsymbol{\beta}_{\mathcal{G}}}\boldsymbol{\beta}_g=(\mathbf{0},\boldsymbol{I}_{p_g},\mathbf{0})\in\mathbb{R}^{p_g\times\sum_{i=1}^{a+b}g_i}$. Combining all $|\mathcal{G}|$ vector equations and rearranging yields the LAR updating direction

$$\frac{d}{dt}\boldsymbol{\beta}_{\mathcal{G}}(t)=-[\boldsymbol{H}_{\mathcal{G}}(t)+\boldsymbol{D}]^{-1}\sqrt{p_g}s'(t)\frac{\boldsymbol{\beta}_g}{\|\boldsymbol{\beta}_g\|}=\frac{s'(t)}{s(t)}[\boldsymbol{H}_{\mathcal{G}}(t)+\boldsymbol{D}]^{-1}\nabla_{\mathcal{G}}f(\boldsymbol{\beta}), \quad (3.22)$$

where $\boldsymbol{D}$ is the block diagonal matrix with blocks

$$\sqrt{p_g}\frac{s(t)}{\|\boldsymbol{\beta}_g\|_2}\left(\boldsymbol{I}_{p_g}-\frac{\boldsymbol{\beta}_g\boldsymbol{\beta}_g^{\mathsf{T}}}{\|\boldsymbol{\beta}_g\|_2^2}\right), \quad g\in\mathcal{G}.$$

2. When $\mathcal{G}_0$ not empty, we assume that $\mathcal{G}_0$ contains a groups, $g_1,\ldots,g_a$, and $\mathcal{G}_1$ contains b groups, $g_{a+1},\ldots,g_{a+b}$. By rearranging the order of groups, we have $\boldsymbol{\beta}_{\mathcal{G}}^{\mathsf{T}}=(\boldsymbol{\beta}_{\mathcal{G}_0}^{\mathsf{T}},\boldsymbol{\beta}_{\mathcal{G}_1}^{\mathsf{T}})$, where $\boldsymbol{\beta}_{\mathcal{G}_0}^{\mathsf{T}}=(\boldsymbol{\beta}_{g_1}^{\mathsf{T}},\boldsymbol{\beta}_{g_2}^{\mathsf{T}},\ldots,\boldsymbol{\beta}_{g_a}^{\mathsf{T}})$ and $\boldsymbol{\beta}_{\mathcal{G}_1}^{\mathsf{T}}=(\boldsymbol{\beta}_{g_{a+1}}^{\mathsf{T}},\boldsymbol{\beta}_{g_{a+2}}^{\mathsf{T}},\ldots,\boldsymbol{\beta}_{g_{a+b}}^{\mathsf{T}})$. For any group $g\in\mathcal{G}_1$, i.e., $\|\boldsymbol{\beta}_g\|_2\neq 0$, the vector equation (3.21) still holds which gives

$$\left[\boldsymbol{H}_{g,\mathcal{G}}(\boldsymbol{\beta})+\frac{\sqrt{p_g}s(t)}{\|\boldsymbol{\beta}_g\|_2}\left(\boldsymbol{I}_{p_g}-\frac{\boldsymbol{\beta}_g\boldsymbol{\beta}_g^{\mathsf{T}}}{\|\boldsymbol{\beta}_g\|_2}\right)D_{\boldsymbol{\beta}_{\mathcal{G}}}\boldsymbol{\beta}_g\right]\frac{d\beta_{\mathcal{G}}(t)}{dt}=\frac{s(t)}{s'(t)}\nabla_g f(\boldsymbol{\beta}). \quad (3.23)$$

Unfortunately it does not hold for any $g\in\mathcal{G}_0$ due to the singularity $\|\boldsymbol{\beta}_g\|_2=0$. First we show that the updating direction of such groups is proportional to their gradient sub-vectors. By the LAR identity (3.20) and the fact $\boldsymbol{\beta}_g(t)=\mathbf{0}_{p_g}$,

$$\frac{[\boldsymbol{\beta}_g(t+\delta t)-\boldsymbol{\beta}_g(t)]/\delta t}{\|[\boldsymbol{\beta}_g(t+\delta t)-\boldsymbol{\beta}_g(t)]/\delta t\|_2}=-\frac{1}{\sqrt{p_g}s(t+\delta t)}\nabla_g f(\boldsymbol{\beta}(t+\delta t))$$

for all $\delta t>0$. Taking limit $\delta t\downarrow 0$ yields

$$\frac{d\boldsymbol{\beta}_g(t)}{dt}=k_g\nabla_g f(\boldsymbol{\beta}), \quad (3.24)$$

where k_g are the constants to be determined.

Equating the norms of the two summand vectors in the LAR identity (3.20) shows $\|\nabla_g f(\boldsymbol{\beta})\|_2^2 = p_g s^2(t)$. Differentiating both sides of this identity with respect to t via chain rule gives

$$d_g f(\boldsymbol{\beta}) \boldsymbol{H}_{g,\mathcal{G}}(\boldsymbol{\beta}) \frac{d\boldsymbol{\beta}_{\mathcal{G}}(t)}{dt} = p_g s(t) s'(t) \tag{3.25}$$

for any $g \in \mathcal{G}_0$. Now substituting (3.24) into the equations (3.23) and (3.25), we obtain

$$\begin{pmatrix} \boldsymbol{A} & \boldsymbol{B}^{\mathsf{T}} \\ \boldsymbol{B} & \boldsymbol{C} \end{pmatrix} \begin{pmatrix} k_1 \\ \vdots \\ k_a \\ \frac{d}{dt}\boldsymbol{\beta}_{\mathcal{G}_1}(t) \end{pmatrix} = \begin{pmatrix} p_{g_1} s'(t) \\ \vdots \\ p_{g_a} s'(t) \\ \nabla_{g_{a+1}} f(\boldsymbol{\beta})/s'(t) \\ \vdots \\ \nabla_{g_{a+b}} f(\boldsymbol{\beta})/s'(t) \end{pmatrix} s(t), \tag{3.26}$$

where $\boldsymbol{A} \in \mathbb{R}^{a \times a}$ has entries $a_{ij} = d_{g_i} f(\boldsymbol{\beta}) \boldsymbol{H}_{g_i,g_j}(\boldsymbol{\beta}) \nabla_{g_j} f(\boldsymbol{\beta})$, $1 \leq i, j \leq a$,

$$\begin{aligned} \boldsymbol{B} &= \begin{bmatrix} \boldsymbol{H}_{\mathcal{G}_1,g_1} \nabla_{g_1} f(\boldsymbol{\beta}), & \cdots, & \boldsymbol{H}_{\mathcal{G}_1,g_a} \nabla_{g_a} f(\boldsymbol{\beta}) \end{bmatrix} \in \mathbb{R}^{\sum_{i=a+1}^{a+b} p_i \times a}, \\ \boldsymbol{C} &= \boldsymbol{H}_{\mathcal{G}_1,\mathcal{G}_1} + \boldsymbol{D}_{\mathcal{G}_1,\mathcal{G}_1} \in \mathbb{R}^{\sum_{i=a+1}^{a+b} p_i \times \sum_{i=a+1}^{a+b} p_i}. \end{aligned}$$

Next we show that the linear system is nonsingular and thus admits a unique solution, when $H_{\mathcal{A}}(\boldsymbol{\beta}(t)) = H_{\mathcal{G},\mathcal{G}}(\boldsymbol{\beta}(t))$ is positive definite. Rewrite

$$\begin{pmatrix} \boldsymbol{A} & \boldsymbol{B}^{\mathsf{T}} \\ \boldsymbol{B} & \boldsymbol{C} \end{pmatrix} = \boldsymbol{E}(\boldsymbol{H}_{\mathcal{A}} + \tilde{\boldsymbol{D}}) \boldsymbol{E}^{\mathsf{T}}$$

where

$$
\begin{aligned}
\boldsymbol{E} &= \begin{pmatrix} d_{g_1} f(\boldsymbol{\beta}) & & & \\ & \ddots & & \\ & & d_{g_a} f(\boldsymbol{\beta}) & \\ & & & \boldsymbol{I}_{\sum_{i=a+1}^{a+b} p_{g_i}} \end{pmatrix} \in \mathbb{R}^{(a+\sum_{i=a+1}^{a+b} p_i)\times \sum_{i=1}^{a+b} p_i}, \\
\tilde{\boldsymbol{D}} &= \begin{pmatrix} \boldsymbol{0}_{\sum_{i=1}^{a} p_{g_i} \times \sum_{i=1}^{a} p_{g_i}} & \\ & \boldsymbol{D}_{\mathcal{G}_1,\mathcal{G}_1} \end{pmatrix} \in \mathbb{R}^{\sum_{i=1}^{a+b} p_i \times \sum_{i=1}^{a+b} p_i}.
\end{aligned}
$$

Combining the facts (i) $\boldsymbol{H}_{\mathcal{A}}$ is positive definite, (ii) $\tilde{\boldsymbol{D}}$ is positive semidefinite, and (iii) $\|\nabla_g f(\boldsymbol{\beta})\|_2^2 = p_g s^2(t) > 0$ for all $g \in \mathcal{G}_0$, $\boldsymbol{E}(\boldsymbol{H} + \tilde{\boldsymbol{D}})\boldsymbol{E}^{\intercal}$ is positive definite. Thus

$$
\begin{pmatrix} k_1 \\ \vdots \\ k_a \\ \frac{d}{dt}\boldsymbol{\beta}_{\mathcal{G}_1}(t) \end{pmatrix} = \begin{pmatrix} \boldsymbol{A} & \boldsymbol{B}^{\intercal} \\ \boldsymbol{B} & \boldsymbol{C} \end{pmatrix}^{-1} \begin{pmatrix} p_{g_1} s'(t) \\ \vdots \\ p_{g_a} s'(t) \\ \nabla_{g_{a+1}} f(\boldsymbol{\beta})/s'(t) \\ \vdots \\ \nabla_{g_{a+b}} f(\boldsymbol{\beta})/s'(t) \end{pmatrix} s(t). \tag{3.27}
$$

Equation (3.27) coupled with (3.24) yields the LAR updating direction for all active predictors.

Chapter 4

Robust Regression for Optimal Individualized Treatment Rules

4.1 Introduction and Motivation

Given the same drug or treatment, different patients may respond quite differently. Factors causing individual variability in drug response are multi-fold and complex. This has raised increasing interests of individualized medicine, where customized medicine or treatment is recommended to each individual according to his/her characteristics, including genetic, physiological, demographic, environmental, and other clinical information. The rule that applied in personalized medicine to match each patient with a target treatment is called individualized treatment rule (ITR), and our goal is to find the "optimal" one, which if followed by the whole patient population would lead to the "best" outcome. For many complex diseases such as cancer and AIDS, the optimal individualized treatment rule or regime is a dynamical treatment process, involving a sequence of treatment decisions made at different time points throughout the disease evolving course.

Q-learning (Watkins and Dayan, 1992; Murphy, 2005) and A-learning (Murphy, 2003; Robins, 2004) are two main approaches for estimating optimal dynamic treatment regimes based on clinical trials or observational data. Q-learning is based on posing a regression model to estimate the conditional expectation of the outcome at each time point, and then applying a backward recursive procedure to fit the model. A-learning, on the other hand, only requires modeling the contrast function of the treatments at each time point, is therefore more flexible and robust to a model misspecification. See Schulte et al. (2012) for a complete review and comparison of these two methods under various scenarios, in terms of the parameter estimation accuracy and the estimation of expected outcomes.

For simplicity, we consider a single stage randomized clinical trial with two treatments. For each patient, the observed data is $(\boldsymbol{X}, A, Y)$, where $\boldsymbol{X} \in \mathcal{X} = \mathbb{R}^p$ denotes the baseline covariates, $A \in \mathcal{A} = \{0, 1\}$ denotes the randomized treatment assigned to the patient, and Y is the real-valued response, which is coded so that higher values indicate more favorable clinical outcomes. An ITR g is a function mapping from $\mathcal{X}$ to $\mathcal{A}$. In Q- and A-learning, the optimal ITR, g_μ^{opt}, is an individualized treatment rule that maximizes the expected outcome $\mathrm{E}[Y^*(g(\boldsymbol{X}))]$. The potential outcomes $Y^*(1)$ and $Y^*(0)$ are the outcomes for an arbitrary individual in population had treatment 1 or 0 been assigned to, respectively. In actuality, at most one of the potential outcomes can be observed for any individual. Define the Q-function $Q(\boldsymbol{x}, a) \triangleq \mathrm{E}(Y|\boldsymbol{x}, a)$. Under some basic assumptions (see Section 4.2 for details), we have

$$g_\mu^{\text{opt}}(\boldsymbol{x}) = \underset{a \in \mathcal{A}}{\operatorname{argmax}}\, Q(\boldsymbol{x}, a). \tag{4.1}$$

We could further define the value function $V_\mu(g) = \mathrm{E}_{\boldsymbol{X}}[Q(\boldsymbol{X}, g(\boldsymbol{X}))]$ which is simply the marginal mean outcome under the ITR g, and $g_\mu^{\text{opt}} = \operatorname{argmax}_g V_\mu(g)$.

All existing methods for optimal treatment regime estimation, including Q-learning and A-learning, belong to mean regression as they estimate the optimal estimator by maximizing expected outcomes. In the case of single decision point, Q-learning is equivalent to the least-squares regression. Least-squares estimates are optimal if the errors are i.i.d. normal random variables. However, skewed, heavy-tailed, heteroscedastic errors or outliers of the response Y

are frequently encountered in clinical trials. In such situations, the efficiency of the least square estimates is impaired. One extreme example is that when the response Y takes i.i.d. Cauchy errors, neither Q-learning nor A-learning can consistently estimate the optimal ITR. It is therefore more reasonable and informative to construct the optimal decision rule based on the conditional median or quantiles of Y given $\boldsymbol{X}$ than based on average effects. Define the τ-th conditional quantile of Y given $(\boldsymbol{X}, A)$ as $Q_\tau(\boldsymbol{X}, A) \triangleq \inf\{y : F_{Y|\boldsymbol{X},A}(y) \geq \tau\}$. Then we define the value function based on the τ-th conditional quantile as $V_{\tau-q}(g) = \mathrm{E}_{\boldsymbol{X}}[Q_\tau(\boldsymbol{X}, g(\boldsymbol{X}))]$, which is an analog to the definition of $V_\mu(g)$. The optimal ITR which maximizes the τ-th conditional quantile is then defined as

$$g_\tau^{\mathrm{opt}}(\boldsymbol{x}) = \underset{a\in\mathcal{A}}{\operatorname{argmax}}\; Q_\tau(\boldsymbol{x}, a),\ \tau \in [0, 1], \tag{4.2}$$

and $g_\tau^{\mathrm{opt}} = \operatorname{argmax}_g V_{\tau-q}(g)$.

In the following, we present a simple example where a quantile-based decision rule is more preferable than a mean-based decision rules. Figure 4.1 plots the conditional density of Y under two treatments, A and B, given a binary covariate X which takes the value of male and female. Under the comparison based on conditional means, A and B are exactly equivalent. However, conditional quantiles provide us more insight. For the male group, the conditional distribution of Y given treatment B is a log-normal and skewed to the right. Therefore, treatment B is less favorable when either 50% or 25% conditional quantile are considered. For the female group, the conditional distribution of Y given treatment A is a standard normal while a Cauchy distribution given treatment B. Therefore, if we make a comparison based on 25% conditional quantile, treatment A is more favorable.

In this chapter, we propose a general framework for optimal individualized treatment rule estimation based on robust regression, including quantile regression and the regression based on Huber's loss and ϵ-insensitive loss. The proposed methodology has the following desired features. First, the new decision rule obtained by maximizing the conditional quantile, which

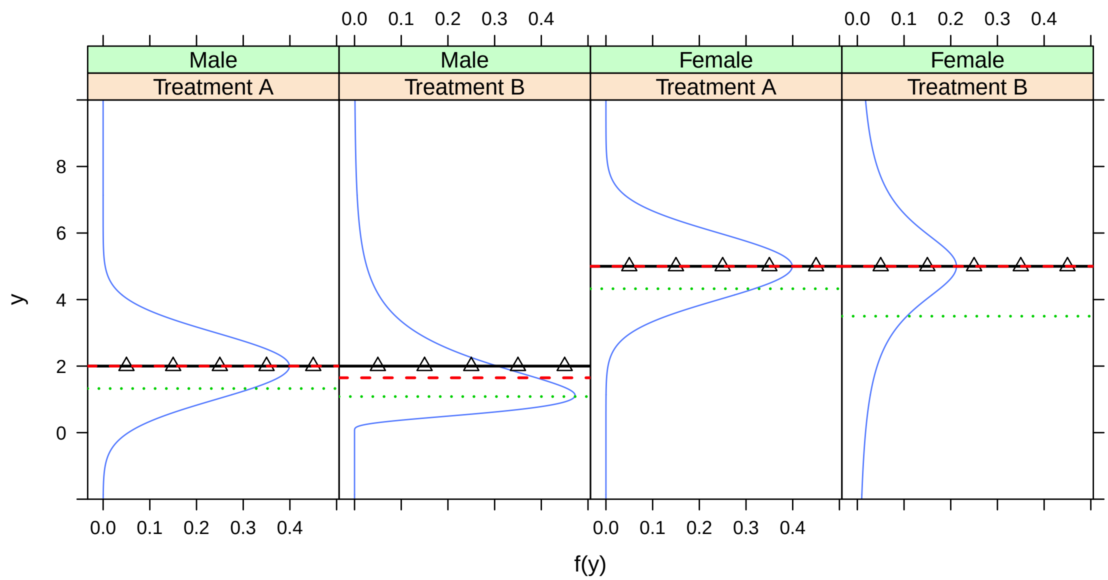


Figure 4.1: The distribution functions of the response Y, in a randomized clinical trial with two treatments, A and B, for male (two panels on the left) and female (two panels on the right). The solid lines with triangle symbol, dashed line, and dotted lines are the conditional mean, 50% quantile, and 25% quantile functions of Y given the gender and the treatment, respectively.

is suitable for skewed, heavy-tailed errors or outliers. To our best knowledge, it is the first quantile regression method for estimating optimal individualized treatment rules. Second, the proposed estimator requires only modeling the contrast function between two treatments, and is therefore robust against misspecification of the baseline function. This property is shared by A-learning. Third, empirical results from our comprehensive numerical study suggest favorable performance of the new robust regression estimator.

The rest of the chapter is organized as follows. In Section 4.2, we first review the classical Q- and A- learning methods. Then we propose the new procedure and method and discuss its connection with existing methods. In Section 4.3, we study and prove the asymptotic properties of the proposed method, including consistency and asymptotic normality. In Section 4.4, a comprehensive numerical study is conducted to assess finite sample performance of the new procedure. In Section 4.5, we apply the method to data from an AIDS study. Concluding remarks are given in Section 4.6. Throughout the chapter, we use upper case letters to denote random variables and lower case letters to denote their values.

4.2 New Optimal Treatment Estimation Framework: Robust Regression

4.2.1 Overview

Recall that, for each patient, we observe $(\boldsymbol{X}, A, Y)$. Here $\boldsymbol{X} \subset \mathbb{R}^p$ is the baseline information, $A \in \mathcal{A} = \{0, 1\}$ is the treatment, and Y is the real-valued response. We first review the potential outcome framework (Neyman, 1923; Rubin, 1974, 1986). The potential outcome $Y^*(a)$ is the outcome for an arbitrary individual has s/he received treatment a. The optimal ITR under mean regression, which maximizes the expected outcome, is $g_\mu^{\text{opt}} = \text{argmax}_{g \in \mathcal{G}} \, \text{E}[Y^*(g(\boldsymbol{X}))]$. Define the propensity score $\pi(\boldsymbol{X}) \triangleq P(A = 1|\boldsymbol{X})$. Following Rubin (1974) and Rubin (1986), we can compute the expectation of the potential outcome under the following two key assumptions.

(C1) **Stable Unit Treatment Value Assumption (SUTVA):** a patient's observed outcome is the same as the potential outcome for the treatment that s/he actually received. Based on Rubin (1986), the SUTVA assumption implies that the value of the potential outcome for a subject does not depend on what treatments other subject receive. Specifically, we can write the SUTVA assumption as

$$Y_i = Y_i^*(1)A_i + Y_i^*(0)(1 - A_i), \ i = 1, \ldots, n.$$

This is also referred as consistency assumption.

(C2) **Strong Ignorability Assumption:** the treatment assignment A for an individual is independent of the potential outcomes conditional on the covariates, i.e., $A \perp \{Y^*(a)\}_{a \in \mathcal{A}} | \boldsymbol{X}$. For a randomized clinical trial, this assumption is satisfied automatically. For an observational study, as clinicians make decisions based only on all past available information, this assumption essentially assumes no unmeasured confounders.

For consistent estimation of the optimal individualized treatment rule, we also need assume

(C3) **Positivity Assumption:** $0 < \pi(\boldsymbol{x}) < 1$, $\forall \boldsymbol{x} \in \mathcal{X}$.

Under assumptions (C1)-(C2), one can show that $g_\mu^{\text{opt}}(\boldsymbol{X}) = \text{argmax}_{a \in \mathcal{A}} \, \text{E}[Y | \boldsymbol{X}, A = a]$. This suggests that, in order to find g_μ^{opt}, we only need to estimate the conditional expectation of Y given $(\boldsymbol{X}, A)$. This result serves as the foundation of Q- and A-learning framework.

Consider the general model $\text{E}(Y | \boldsymbol{X}, A) = h_0(\boldsymbol{X}) + AC_0(\boldsymbol{X})$, where $h_0(\boldsymbol{X})$ represents the baseline effect, and $C_0(\boldsymbol{X})$ denotes the contrast effect as

$$C_0(\boldsymbol{X}) = \text{E}(Y | \boldsymbol{X}, A = 1) - \text{E}(Y | \boldsymbol{X}, A = 0).$$

Therefore, $g_\mu^{\text{opt}}(\boldsymbol{X}) = 1\{C_0(\boldsymbol{X}) > 0\}$. In Q-learning, a parametric model is often employed as a working model,

$$\text{E}(Y | \boldsymbol{X}, A) = h(\boldsymbol{X}; \boldsymbol{\gamma}) + AC(\boldsymbol{X}; \boldsymbol{\beta}), \tag{4.3}$$

where $h(\boldsymbol{X};\boldsymbol{\gamma})$ and $C(\boldsymbol{X};\boldsymbol{\beta})$ are posited parametric models for $h_0(\boldsymbol{X})$ and $C_0(\boldsymbol{X})$ respectively. Commonly a linear model is assumed for simplicity and interpretability, i.e., $h(\boldsymbol{X};\boldsymbol{\gamma}) = \boldsymbol{\gamma}^\intercal \tilde{\boldsymbol{X}}$ and $C(\boldsymbol{X};\boldsymbol{\beta}) = \boldsymbol{\beta}^\intercal \tilde{\boldsymbol{X}}$, where $\tilde{\boldsymbol{X}} = (\mathbf{1}, \boldsymbol{X}^\intercal)^\intercal$. Given the observation $\{(Y_i, \boldsymbol{X}_i, A_i);\ i = 1, \ldots, n\}$, the Q-learning procedure estimates the parameters $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ by minimizing the squared error loss

$$L_{1n}(\boldsymbol{\beta},\boldsymbol{\gamma}) = \frac{1}{n}\sum_{i=1}^{n}\left[Y_i - h(\boldsymbol{X}_i;\boldsymbol{\gamma}) - A_i C(\boldsymbol{X}_i;\boldsymbol{\beta})\right]^2. \tag{4.4}$$

Denote the optimized point as $(\hat{\boldsymbol{\beta}}^Q, \hat{\boldsymbol{\gamma}}^Q)$. The estimated optimal ITR based on Q-learning is then $\hat{g}^Q(\boldsymbol{x}) \triangleq 1\{C(\boldsymbol{x};\hat{\boldsymbol{\beta}}^Q) > 0\}$, which is a consistent estimator of $g_\mu^{\mathrm{opt}}(\boldsymbol{x})$ if both $h(\boldsymbol{X};\boldsymbol{\gamma})$ and $C(\boldsymbol{X};\boldsymbol{\beta})$ are correctly specified.

A-learning is a semiparametric improvement of Q-learning by modeling only the contrast function $C_0(\boldsymbol{X})$ rather than the full Q-function. This is reasonable based on the observation that the optimal ITR g_μ^{opt} only depends on $C_0(\boldsymbol{X})$. By positing $C(\boldsymbol{X};\boldsymbol{\beta})$ for the contrast function, in A-learning, one can estimate coefficients $\boldsymbol{\beta}$ by solving the following estimating equation

$$\sum_{i=1}^{n}\lambda(\boldsymbol{X}_i)\left(A_i - \pi(\boldsymbol{X}_i)\right)\left[Y_i - A_i C(\boldsymbol{X}_i;\boldsymbol{\beta}) - h(\boldsymbol{X}_i)\right] = 0, \tag{4.5}$$

where $\lambda(\boldsymbol{X}_i)$ and $h(\boldsymbol{X}_i)$ are arbitrary functions, and $\lambda(\boldsymbol{X}_i)$ has the same dimension as $\boldsymbol{\beta}$. Denote the solution to (4.5) by $\hat{\boldsymbol{\beta}}^A$. If $\mathrm{var}(Y|X)$ is constant and $C(\boldsymbol{X}_i;\boldsymbol{\beta})$ is correctly specified, the optimal choices of $\lambda(\cdot)$ and $h(\cdot)$ are $\lambda(\boldsymbol{X}_i;\boldsymbol{\beta}) = \partial/\partial\boldsymbol{\beta} C(\boldsymbol{X}_i;\boldsymbol{\beta})$ and $h(\boldsymbol{X}_i) = h_0(\boldsymbol{X}_i)$ (Robins, 2004). In practice, one may pose models, say $\pi(\boldsymbol{X}_i;\boldsymbol{\phi})$ and $h(\boldsymbol{X}_i;\boldsymbol{\gamma})$ for $\pi(\boldsymbol{X}_i)$ and $h(\boldsymbol{X}_i)$ respectively, and take $\lambda(\boldsymbol{X}_i;\boldsymbol{\beta}) = \partial/\partial\boldsymbol{\beta} C(\boldsymbol{X}_i;\boldsymbol{\beta})$. Under randomized designs, the propensity score $\pi(\boldsymbol{X}_i)$ is known. Otherwise, a logistic model can be proposed. Under the assumption that $C(\boldsymbol{X};\boldsymbol{\beta})$ is correctly specified, the double robustness property of A-learning states that as long as one of $\pi(\boldsymbol{X};\boldsymbol{\phi})$ and $h(\boldsymbol{X};\boldsymbol{\gamma})$ is correctly specified, $\hat{g}^A(\boldsymbol{x}) \triangleq 1\{C(\boldsymbol{x};\hat{\boldsymbol{\beta}}^A) > 0\}$ is consistent estimator of $g_\mu^{\mathrm{opt}}(\boldsymbol{x})$.

Recently, Lu et al. (2013) propose a variant of A-learning by a loss-based learning framework. Rewrite

$$
\begin{aligned}
\mathrm{E}(Y|\boldsymbol{X}, A) =& h_0(\boldsymbol{X}) + AC_0(\boldsymbol{X}) \\
=& \varphi_0(\boldsymbol{X}) + (A - \pi(\boldsymbol{X}))C_0(\boldsymbol{X}),
\end{aligned}
$$

where $\varphi_0(\boldsymbol{X}) = h_0(\boldsymbol{X}) + \pi(\boldsymbol{X})C_0(\boldsymbol{X})$. Based on the expression above, Lu et al. (2013) propose to estimate $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ by minimizing the following loss function

$$
L_{2n}(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \frac{1}{n}\sum_{i=1}^{n}\left[Y_i - \varphi(\boldsymbol{X}_i; \boldsymbol{\gamma}) - (A_i - \pi(\boldsymbol{X}_i))C(\boldsymbol{X}_i; \boldsymbol{\beta})\right]^2, \tag{4.6}
$$

where $\varphi(\boldsymbol{X}; \boldsymbol{\gamma})$, $C(\boldsymbol{X}; \boldsymbol{\beta})$ are proposed models for $\varphi_0(\boldsymbol{X})$ and $C_0(\boldsymbol{X})$ respectively. Denote the minimizer of (4.6) as $(\hat{\boldsymbol{\beta}}^A_{LS}, \hat{\boldsymbol{\gamma}}^A_{LS})$. Lu et al. (2013) show that $\hat{g}^A_{LS}(\boldsymbol{x}) \triangleq 1\{C(\boldsymbol{x}; \hat{\boldsymbol{\beta}}^A_{LS}) > 0\}$ is consistent estimator of $g^{\mathrm{opt}}_{\mu}(\boldsymbol{x})$ when the propensity score $\pi(\boldsymbol{X})$ is known or can be consistently estimated from the data, and $C(\boldsymbol{X}; \boldsymbol{\beta})$ is correctly specified. We refer to this method as least square A-learning (lsA-learning).

One main advantage of the lsA-learning, compared to the classical A-learning, is its square loss, making the procedure easy to be coupled with penalized regression to achieve variable selection in high dimensional data. Specifically, Lu et al. (2013) propose to identify important nonzero coefficients in $\boldsymbol{\beta}$ by applying an adaptive LASSO penalty to (4.6). Under some regularity conditions, both the selection consistency and asymptotic normality of the estimator are established in Lu et al. (2013). The downside of lsA-learning is that one direction of the double robustness property of the classical A-learning is lost, i.e., when $\varphi(\boldsymbol{X}; \boldsymbol{\gamma})$ is correctly specified, $\boldsymbol{\beta}$ may still not be consistent if the propensity score $\pi(\boldsymbol{X})$ is not consistently estimated. Finally, it can be shown that lsA-learning and Q-learning are equivalent when $\pi(\boldsymbol{X})$ is constant and both $\varphi(\boldsymbol{X}; \boldsymbol{\gamma})$ and $C(\boldsymbol{X}; \boldsymbol{\beta})$ take the linear form (with the space of $C(\boldsymbol{X}; \boldsymbol{\beta})$ included in the space of $\varphi(\boldsymbol{X}; \boldsymbol{\gamma})$). Similar properties hold for A-learning (Chakraborty et al., 2010; Schulte et al., 2012).

4.2.2 New Proposal: Robust Regression

Skewed, heavy-tailed, heteroscedastic errors or outliers of the response Y are frequently encountered in clinical trials. It is well known that ordinary least square estimation fails to produce a reliable estimator in such situations. The immediate consequence is the efficiency loss in the estimators produced by Q-, A-, and lsA-learning. This motivates us to adopt robust regression techniques in optimal treatment regime estimation.

We consider the following additive model,

$$Y_i = \varphi_0(\boldsymbol{X}_i) + (A_i - \pi(\boldsymbol{X}_i))C(\boldsymbol{X}_i; \beta_0) + \epsilon_i, \; i = 1, \ldots, n, \tag{4.7}$$

where $\varphi_0(\boldsymbol{X})$ is the baseline function, $C(\boldsymbol{X}; \beta_0)$ is the contrast function, $\pi(\boldsymbol{X})$ is the propensity score, and ϵ is the error term which satisfies the conditional independence assumption $\epsilon \perp A|\boldsymbol{X}$. We point out that the error term defined in (4.7) can be very general. For example, we could take $\epsilon = \sum_{j=1}^{K} \sigma_j(\boldsymbol{X})e_j$ for any $K \geq 1$ that allows the error distribution to change with $\boldsymbol{X}$, used to model heterogeneous errors, where $\sigma_j(\boldsymbol{X})$ are arbitrary positive functions and $e_j \perp (A, \boldsymbol{X})$ for all $j = 1, \ldots, K$. Throughout the chapter, we assume $\{(Y_i, \boldsymbol{X}_i, A_i, \epsilon_i), i = 1, \ldots, n\}$ are i.i.d random samples of the population.

We propose to estimate $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ by minimizing

$$L_{3n}(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \frac{1}{n} \sum_{i=1}^{n} M\left[Y_i - \varphi(\boldsymbol{X}_i; \boldsymbol{\gamma}) - (A_i - \pi(\boldsymbol{X}_i))C(\boldsymbol{X}_i; \boldsymbol{\beta})\right], \tag{4.8}$$

where $\boldsymbol{\gamma} \in \Gamma$, $\boldsymbol{\beta} \in \mathcal{B}$ and $M : \mathbb{R}^{\rightarrow}[0, \infty)$ is a convex function with minimum achieved at 0. Denote the minimizer of (4.8) as $(\hat{\boldsymbol{\beta}}_M^R, \hat{\boldsymbol{\gamma}}_M^R)$, and the estimated ITR is then $\hat{g}_M^R(\boldsymbol{x}) \triangleq 1\{C(\boldsymbol{x}; \hat{\boldsymbol{\beta}}_M^R) > 0\}$. In the following, we refer the robust regression with loss function $M(x)$ as RR(M)-learning. In this thesis, we consider the following three types of loss functions, i.e., the pinball loss

$$M(x) = \rho_\tau(x) \triangleq \begin{cases} (\tau - 1)x, & \text{if } x < 0 \\ \tau x, & \text{if } x \geq 0 \end{cases} \tag{4.9}$$

where $0 < \tau < 1$, the Huber loss

$$M(x) = H_\alpha(x) \triangleq \begin{cases} 0.5x^2, & \text{if } |x| < \alpha \\ \alpha|x| - 0.5\alpha^2, & \text{if } |x| \geq \alpha \end{cases} \tag{4.10}$$

for some $\alpha > 0$, and the ϵ-insensitive loss

$$M(x) = J_\epsilon(x) \triangleq \max\{0, |x| - \epsilon\} \tag{4.11}$$

for some $\epsilon > 0$. The pinball loss are frequently applied for quantile regression (Koenker, 2005), and the Huber losses and the ϵ-insensitive are robust against heavy tailed errors or outliers. Plots of these loss functions are given in Figure 4.2 together with the squared error loss. A dramatic difference of pinball loss, Huber loss and ϵ-insensitive loss, compared with the square loss, is that they penalize large deviances linearly instead of quadratically. This property makes them more robust when dealing with responses with non-normal type of errors.

4.3 Asymptotic Properties

4.3.1 Consistency of Robust Regression: Pinball Loss

Under the conditional independence assumption $\epsilon \perp A|\boldsymbol{X}$, we have

$$\begin{aligned} Q(\boldsymbol{X}, A) =& \varphi_0(\boldsymbol{X}) + (A - \pi(\boldsymbol{X}))C(\boldsymbol{X}; \beta_0) + \mu_\epsilon(\boldsymbol{X}); \\ Q_\tau(\boldsymbol{X}, A) =& \varphi_0(\boldsymbol{X}) + (A - \pi(\boldsymbol{X}))C(\boldsymbol{X}; \beta_0) + F_\epsilon^{-1}(\boldsymbol{X}; \tau). \end{aligned}$$

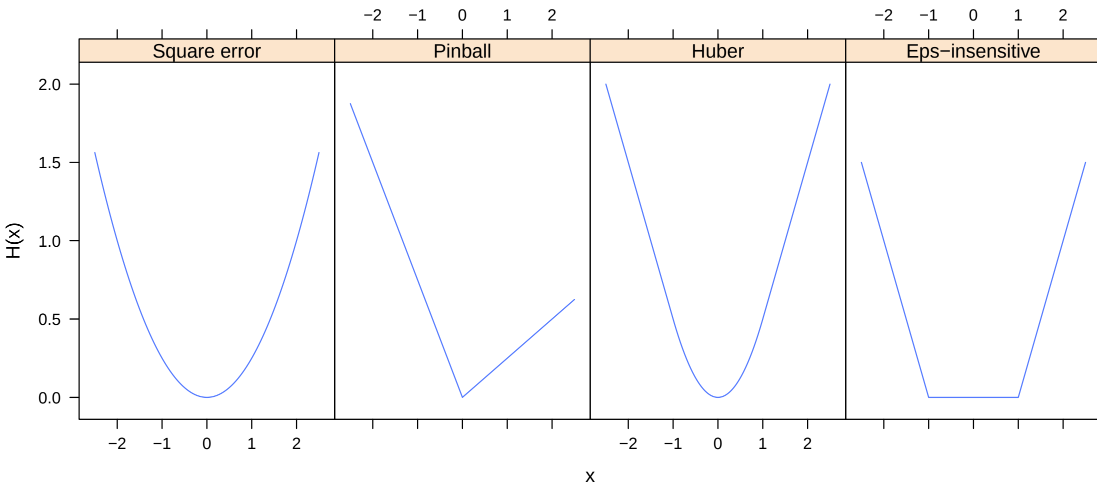


Figure 4.2: Various loss functions: squared error loss, pinball loss with $\tau = 0.25$, Huber loss with $\alpha = 1$ and ϵ-insensitive loss with $\epsilon = 1$.

where $\mu_\epsilon(\boldsymbol{X})$ and $F_\epsilon^{-1}(\boldsymbol{X};\tau)$ denote the mean and the τ-th quantile of ϵ conditional on $\boldsymbol{X}$ respectively. Therefore, in this situation, we have $g_\mu^{\text{opt}} = g_\tau^{\text{opt}} = 1\{C(\boldsymbol{X};\boldsymbol{\beta}_0) > 0\}$. In other words, the underlying ITR which maximize the population mean and τ-th quantile are equivalent. For a good ITR $\hat{g} = 1\{C(\boldsymbol{X};\hat{\boldsymbol{\beta}}) > 0\}$, it is reasonable to require $\hat{\boldsymbol{\beta}}$ to be a consistent estimator of $\boldsymbol{\beta}_0$. This consistency result is first shown for the robust regression with pinball loss, which is given in Theorem 4.1. We allocate all the proofs into the Supplement.

Theorem 4.1. *Under regularity conditions (A1)-(A8) in the Supplement, if the contrast function in* (4.7) *is correctly specified and* $\pi(\boldsymbol{x})$ *is known, then* $\hat{\boldsymbol{\beta}}^R_{\rho(\tau)} \xrightarrow{p} \boldsymbol{\beta}_0$ *for all* $\tau \in (0,1)$*, where* $\hat{\boldsymbol{\beta}}^R_{\rho(\tau)}$ *is the solution of* (4.8) *when* $M(x) = \rho_\tau(x)$.

Remarks:

1. Theorem 4.1 doesn't assume the finiteness of $E(Y)$. Therefore it can be applied to the cases when ϵ_i follows a Cauchy distribution.

2. After fitting the model, the Assumption (A2), $\epsilon \perp A|\boldsymbol{X}$, can be verified by applying conditional independence test with $\hat{r}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}})$ and A given $\boldsymbol{X}$, where $\hat{r}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}})$ is the residual

and can be computed as $\hat{r}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) = Y - \varphi(\boldsymbol{X}; \hat{\boldsymbol{\gamma}}) - (A - \pi(\boldsymbol{X}))C(\boldsymbol{X}; \hat{\boldsymbol{\beta}})$. See Lawrance (1976); Su and White (2007); Song (2009); Huang (2010); Zhang et al. (2012) for more discussion of conditional independence hypothesis tests. In particular, we demonstrate the usefulness of the test by applying the Kernel-based conditional independence test (KCI-test, Zhang et al. (2012)) in Section 4.5. KCI-test doesn't assume functional forms among variables and thus suits our need.

When the conditional independence assumption ($\epsilon \perp A|\boldsymbol{X}$) does not hold, $\hat{\boldsymbol{\beta}}^R_{\rho(\tau)}$ may no longer be a consistent estimator of $\boldsymbol{\beta}_0$. This is intuitively reasonable as ϵ contains extra information with respect to A. In fact, a general result which can be derived in this case is that, $(\hat{\boldsymbol{\beta}}^R_{\rho(\tau)}, \hat{\boldsymbol{\gamma}}^R_{\rho(\tau)})$ minimizes a weighed mean-square error loss function with specification error (Angrist et al., 2006; Lee, 2013).

Instead of assuming response Y takes an additive error term ϵ as in (4.7), we assume the conditional quantile function $Q_\tau(\boldsymbol{X}, A) = \varphi_0(\boldsymbol{X}) + (A - \pi(\boldsymbol{X}))C(\boldsymbol{X}; \boldsymbol{\beta}_0(\tau))$, where we redundantly represent the baseline function and contrast function as $\varphi_0(\cdot)$ and $C(\cdot)$ respectively. Notice that we use $\boldsymbol{\beta}_0(\tau)$ instead of $\boldsymbol{\beta}_0$ to emphasize that the true $\boldsymbol{\beta}$ may vary with respect to τ. The proposed model is $\hat{Q}(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \varphi(\boldsymbol{X}; \boldsymbol{\gamma}) + (A - \pi(\boldsymbol{X}))C(\boldsymbol{X}; \boldsymbol{\beta})$ with $C(\boldsymbol{X}; \boldsymbol{\beta})$ correctly specified. Define

$$(\boldsymbol{\beta}(\tau), \boldsymbol{\gamma}(\tau)) = \underset{\boldsymbol{\beta} \in \mathcal{B}, \boldsymbol{\gamma} \in \Gamma}{\operatorname{argmin}} \mathrm{E}\left\{\rho_\tau[Y - \hat{Q}(\boldsymbol{\beta}, \boldsymbol{\gamma})] - \rho_\tau[Y - \hat{Q}(\boldsymbol{\beta}', \boldsymbol{\gamma}')]\right\} \tag{4.12}$$

where $(\boldsymbol{\beta}', \boldsymbol{\gamma}')$ is any fixed point in $\mathcal{B} \times \Gamma$. Define the QR specification error as $\Delta_\tau(\boldsymbol{X}, A; \boldsymbol{\beta}, \boldsymbol{\gamma}) \triangleq \hat{Q}(\boldsymbol{\beta}, \boldsymbol{\gamma}) - Q_\tau(\boldsymbol{X}, A)$. Define the quantile-specific residual as $\epsilon_\tau \triangleq Y - Q_\tau(\boldsymbol{X}, A)$ with conditional density function $f_{\epsilon_\tau}(\cdot|\boldsymbol{X}, A)$. Then we have the following approximation theorem. The proof of the theorem follows Theorem 1 of Angrist et al. (2006), and is omitted for brevity.

Theorem 4.2. *Suppose that (i) the conditional density $f_Y(y|\boldsymbol{X}, A)$ exists a.s.; (ii)$E[Q_\tau(\boldsymbol{X}, A)]$*

and $E[\Delta_\tau^2(\boldsymbol{X}, A; \boldsymbol{\beta}, \boldsymbol{\gamma})]$ *are finite; (iii)* $(\boldsymbol{\beta}(\tau), \boldsymbol{\gamma}(\tau))$ *uniquely solves* (4.12). *Then*

$$(\boldsymbol{\beta}(\tau), \boldsymbol{\gamma}(\tau)) = \underset{\boldsymbol{\beta}, \boldsymbol{\gamma}}{\operatorname{argmin}}\, E[w_\tau(\boldsymbol{X}, A; \boldsymbol{\beta}, \boldsymbol{\gamma})\Delta_\tau^2(\boldsymbol{X}, A; \boldsymbol{\beta}, \boldsymbol{\gamma})] \tag{4.13}$$

where

$$w_\tau(\boldsymbol{X}, A; \boldsymbol{\beta}, \boldsymbol{\gamma}) = \int_0^1 (1-u) f_{\epsilon_\tau}(u\Delta_\tau(\boldsymbol{X}, A; \boldsymbol{\beta}, \boldsymbol{\gamma})|\boldsymbol{X}, A)du. \tag{4.14}$$

Remarks:

1. Theorem 4.2 shows that $\hat{Q}\,(\boldsymbol{\beta}(\tau), \boldsymbol{\gamma}(\tau))$ is a weighted least square approximation to $Q_\tau(\boldsymbol{X}, A)$. In other word, $\varphi(\boldsymbol{X}; \boldsymbol{\gamma}(\tau)) + (A - \pi(\boldsymbol{X}))C(\boldsymbol{X}; \boldsymbol{\beta}(\tau))$ is close to $\varphi_0(\boldsymbol{X}) + (A - \pi(\boldsymbol{X}))C(\boldsymbol{X}; \boldsymbol{\beta}_0(\tau))$. So even though it is not true that $\boldsymbol{\beta}(\tau) = \boldsymbol{\beta}_0(\tau)$ holds exactly, the difference between them is small in general . This coupled with the fact that $\hat{\boldsymbol{\beta}}^R_{\rho(\tau)} \xrightarrow{p} \boldsymbol{\beta}(\tau)$ (proved in Theorem 4.4), leads to the conclusion that approximately ITR $\hat{g}^R_{\rho(\tau)}(\boldsymbol{x})$ ($\triangleq 1\{C(\boldsymbol{x}; \hat{\boldsymbol{\beta}}^R_{\rho(\tau)}) > 0\}$) maximizes the τ-th conditional quantile. This observation is justified numerically in Section 4.4.

2. When there exists $\boldsymbol{\gamma}_0 \in \Gamma$ such that $\varphi_0(\boldsymbol{X}) \equiv \varphi(\boldsymbol{X}; \boldsymbol{\gamma}_0)$, then we have $\boldsymbol{\beta}(\tau) = \boldsymbol{\beta}_0(\tau)$.

4.3.2 Consistency of Robust Regression: Other Losses

Under model (4.7) and the assumption $\epsilon \perp A|\boldsymbol{X}$, similar consistency results can be established for Huber loss and the ϵ-insensitive loss, as stated in Theorem 4.3.

Theorem 4.3. *Under regularity conditions (A1)-(A8), if the contrast function in* (4.7) *is correctly specified and* $\pi(\boldsymbol{x})$ *is known, then we have*

(a) $\hat{\boldsymbol{\beta}}^R_{H(\alpha)} \xrightarrow{p} \boldsymbol{\beta}_0$ *for all* $\alpha > 0$*, where* $\hat{\boldsymbol{\beta}}^R_{H(\alpha)}$ *is the solution of* (4.8) *when* $M(x) = H_\alpha(x)$*;*

(b) $\hat{\boldsymbol{\beta}}^R_{J(\epsilon)} \xrightarrow{p} \boldsymbol{\beta}_0$ *for all* $\epsilon > 0$*, where* $\hat{\boldsymbol{\beta}}^R_{J(\epsilon)}$ *is the solution of* (4.8) *when* $M(x) = J_\epsilon(x)$*.*

4.3.3 Asymptotic Normality: Pinball Loss

Without loss of generality, in this section we assume both the $\varphi(\boldsymbol{X};\boldsymbol{\gamma})$ and $C(\boldsymbol{X};\boldsymbol{\beta})$ take the linear form: $\varphi(\boldsymbol{X};\boldsymbol{\gamma}) = \tilde{\boldsymbol{X}}^\intercal\boldsymbol{\gamma}$ and $C(\boldsymbol{X};\boldsymbol{\beta}) = \tilde{\boldsymbol{X}}^\intercal\boldsymbol{\beta}$, where $\tilde{\boldsymbol{X}} = (1, \boldsymbol{X}^\intercal)^\intercal$. Denote $\hat{\boldsymbol{\beta}}(\tau) = \hat{\boldsymbol{\beta}}^R_{\rho(\tau)}$ and $\hat{\boldsymbol{\gamma}}(\tau) = \hat{\boldsymbol{\gamma}}^R_{\rho(\tau)}$. Denote $\boldsymbol{W} = ([A-\pi(\boldsymbol{X})]\tilde{\boldsymbol{X}}^\intercal, \tilde{\boldsymbol{X}}^\intercal)^\intercal$, $\boldsymbol{\theta}(\tau) = (\boldsymbol{\beta}(\tau)^\intercal, \boldsymbol{\gamma}(\tau)^\intercal)^\intercal$, $\hat{\boldsymbol{\theta}}(\tau) = (\hat{\boldsymbol{\beta}}(\tau)^\intercal, \hat{\boldsymbol{\gamma}}(\tau)^\intercal)^\intercal$ and $J(\tau) \triangleq \mathrm{E}\left[f_Y(\boldsymbol{W}^\intercal\boldsymbol{\theta}(\tau)|\boldsymbol{X}, A)\boldsymbol{W}\boldsymbol{W}^\intercal\right]$. Under the following regularity conditions, which is the same as the assumptions assumed in Angrist et al. (2006) and Lee (2013), we have the asymptotic normality of $\hat{\boldsymbol{\theta}}(\tau)$, which is given in Theorem 4.4.

(B1) $\{(Y_i, \boldsymbol{X}_i, A_i, \epsilon_i), i = 1, \ldots, n\}$ are i.i.d random variables;

(B2) the conditional density $f_Y(y|\boldsymbol{X} = \boldsymbol{x}, A = a))$ exists, and is bounded and uniformly continuous in y, uniformly in $\boldsymbol{x}$ over the support of $\boldsymbol{X}$;

(B3) $J(\tau)$ is positive definite for all $\tau \in (0, 1)$, where $\boldsymbol{\theta}(\tau)$ is uniquely defined in (4.12);

(B4) $\mathrm{E}\|\boldsymbol{X}\|^{2+\epsilon}$ for some $\epsilon > 0$.

Theorem 4.4. *If regularity conditions (B1)-(B4) are hold, we have*

1. ***(Uniform Consistency)*** $\sup_\tau \|\hat{\boldsymbol{\theta}}(\tau) - \boldsymbol{\theta}(\tau)\| = o_p(1)$*;*

2. ***(Asymptotic Normality)*** $J(\cdot)\sqrt{n}(\hat{\boldsymbol{\theta}}(\cdot) - \boldsymbol{\theta}(\cdot))$ *converge in distribution to a zero mean Gaussian process with covariance function* $\Sigma(\tau, \tau')$ *defined as*

$$\Sigma(\tau, \tau') = E\left[(\tau - 1\{Y < \boldsymbol{W}^\intercal\boldsymbol{\theta}(\tau)\})\left(\tau' - 1\{Y < \boldsymbol{W}^\intercal\boldsymbol{\theta}(\tau)\}\right)\boldsymbol{W}\boldsymbol{W}^\intercal\right]. \tag{4.15}$$

The proof is given in Angrist et al. (2006), and the asymptotic covariance matrix of $\hat{\boldsymbol{\theta}}(\tau)$ can be estimated by either a bootstrap procedure (Hahn, 1997) or a nonparametric kernel method (Angrist et al., 2006). We adopt the parametric bootstrap approach to estimate the asymptotic covariance matrix in Section 4.5. Under model (4.7) the result of Theorem 4.4 can be further simplified, which is given in Theorem 4.5.

Theorem 4.5. *Under the condition of Theorem 4.4, if further we assume* $Y = \varphi_0(\boldsymbol{X}) + (A - \pi(\boldsymbol{X}))\tilde{\boldsymbol{X}}^\intercal\boldsymbol{\beta}_0 + \epsilon$, *and* $\epsilon \perp A|\boldsymbol{X}$, *then*

1. $\sup_\tau \|\hat{\boldsymbol{\beta}}(\tau) - \boldsymbol{\beta}_0\| = o_p(1)$;

2. $\sqrt{n}(\hat{\boldsymbol{\beta}}(\tau) - \boldsymbol{\beta}_0) \xrightarrow{d} N(\mathbf{0}, J_{11}^{-1}(\tau)\Sigma_{11}(\tau,\tau)J_{11}^{-1}(\tau))$, *where*

$$\begin{aligned} J_{11}(\tau) =& E\left[f_\epsilon\left(\tilde{\boldsymbol{X}}^\intercal\boldsymbol{\gamma}(\tau) - \varphi_0(\boldsymbol{X})|\boldsymbol{X}\right)\pi(\boldsymbol{X})(1-\pi(\boldsymbol{X}))\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\intercal\right], \\ \Sigma_{11}(\tau,\tau) =& E\left\{\left[\tau - 1\{\epsilon < \tilde{\boldsymbol{X}}^\intercal\boldsymbol{\gamma}(\tau) - \varphi_0(\boldsymbol{X})\}\right]^2 \pi(\boldsymbol{X})(1-\pi(\boldsymbol{X}))\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\intercal\right\}. \end{aligned}$$

Furthermore, we have $\Sigma_{11}(\tau,\tau) \leq \left(\tau^2 + |1-2\tau|\right) E\left[\pi(\boldsymbol{X})(1-\pi(\boldsymbol{X}))\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\intercal\right]$.

Comparing the asymptotic normality of $\hat{\boldsymbol{\beta}}(\tau)$ with $\hat{\boldsymbol{\beta}}_{LS}^A$ yields interesting insights. Assuming that $\mathrm{E}(Y|\boldsymbol{X}, A) = \varphi_0(\boldsymbol{X}) + (A - \pi(\boldsymbol{X}))\tilde{\boldsymbol{X}}^\intercal\boldsymbol{\beta}_0$ and $(\boldsymbol{\beta}_0, \boldsymbol{\gamma}^*) = \mathrm{argmin}_{(\boldsymbol{\beta},\boldsymbol{\gamma})}\ \mathrm{E}[Y - \varphi(\boldsymbol{X};\boldsymbol{\gamma}) - (A - \pi(\boldsymbol{X}))\tilde{\boldsymbol{X}}^\intercal\boldsymbol{\beta}]^2$, the asymptotic normality property of $\hat{\boldsymbol{\beta}}_{LS}^A$ can then be established, which is summarized in Theorem 4.6. Its proof has been omitted, and readers are referred to Lu et al. (2013).

Theorem 4.6. *Under the regularity condition of A1-A4 of Lu et al. (2013),*

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{LS}^A - \boldsymbol{\beta}_0) \xrightarrow{d} N(0, U_{11}^{-1}\Omega_{11}U_{11}^{-1}), \tag{4.16}$$

where

$$U_{11} = E\left[\pi(\boldsymbol{X})(1-\pi(\boldsymbol{X}))\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\intercal\right]$$

and

$$\Omega_{11} = E\left[\{\varphi_0(\boldsymbol{X}) - \varphi(\boldsymbol{X};\boldsymbol{\gamma}^*) + \epsilon\}^2 \pi(\boldsymbol{X})(1-\pi(\boldsymbol{X}))\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\intercal\right].$$

Remarks:

1. When the family of functions $\{\varphi(\boldsymbol{X};\boldsymbol{\gamma}), \boldsymbol{\gamma} \in \Gamma\}$ cannot well approximate the unknown baseline function $\varphi_0(\boldsymbol{X})$, the Ω_{11} term in the asymptotic variance of $\hat{\boldsymbol{\beta}}_{LS}^A$ may explode, which makes $\hat{\boldsymbol{\beta}}_{LS}^A$ less efficient than $\hat{\boldsymbol{\beta}}(\tau)$.

2. When $Y = \tilde{\boldsymbol{X}}^\intercal \gamma_0 + (A - \pi(\boldsymbol{X}))\tilde{\boldsymbol{X}}^\intercal \boldsymbol{\beta}_0 + \epsilon$, $\epsilon \perp (A, \boldsymbol{X})$, $\pi(\boldsymbol{X}) \equiv 0.5$ and $\epsilon \sim N(0, \sigma^2)$, the asymptotic variance of $\hat{\boldsymbol{\beta}}(\tau = 0.5)$ is $2\pi\sigma^2 \mathrm{E}(\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\intercal)^{-1}$, which is strictly larger than $4\sigma^2 \mathrm{E}(\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\intercal)^{-1}$ (the asymptotic variance of $\hat{\boldsymbol{\beta}}_{LS}^{A}$).

4.4 Numerical Results: Simulation Studies

To demonstrate finite sample performance of the proposed robust regression methods for optimal treatment rule estimation, we conduct two simulation studies: the errors independent with treatments, and the errors interactive with treatments, respectively.

4.4.1 Simulation Study I: error terms independent with treatment

We consider the following two models with p=3,

- Model I:

$$Y_i = 1 + (X_{i1} - X_{i2})(X_{i1} + X_{i3}) + [A_i - \pi(\boldsymbol{X}_i)]\boldsymbol{\beta}_0^\intercal \tilde{\boldsymbol{X}}_i + \sigma(\boldsymbol{X}_i)\epsilon_i,$$

  where $\boldsymbol{X}_i = (X_{i1}, X_{i2}, X_{i3})^\intercal$ are multivariate normal with mean 0, variance 1. $\tilde{\boldsymbol{X}}_i = (1, \boldsymbol{X}_i^\intercal)^\intercal$, $\mathrm{Corr}(X_{ij}, X_{ik}) = 0.5^{|j-k|}$ and $\boldsymbol{\beta}_0 = (0, 1, -1, 1)^\intercal$.

- Model II:

$$Y_i = \boldsymbol{\gamma}_0^\intercal \tilde{\boldsymbol{X}}_i + [A_i - \pi(\boldsymbol{X}_i)]\boldsymbol{\beta}_0^\intercal \tilde{\boldsymbol{X}}_i + \sigma(\boldsymbol{X}_i)\epsilon_i,$$

  where $\boldsymbol{\gamma}_0^\intercal = (0.5, 4, 1, -3)$, and $\boldsymbol{X}_i$, $\tilde{\boldsymbol{X}}_i$ and $\boldsymbol{\beta}_0$ are the same as Model I.

We take linear forms for both the baseline and the contrast functions, where $\varphi(\boldsymbol{X};\boldsymbol{\gamma}) = \boldsymbol{\gamma}^\intercal \tilde{\boldsymbol{X}}$ and $C(\boldsymbol{X};\boldsymbol{\beta}) = \boldsymbol{\beta}^\intercal \tilde{\boldsymbol{X}}$. We assume the propensity scores $\pi(\cdot)$ are known, and we study both the constant case ($\pi(\boldsymbol{X}_i) = 0.5$) and the non-constant case ($\pi(\boldsymbol{X}_i) = \mathrm{logit}(\boldsymbol{X}_{i1} - \boldsymbol{X}_{i2})$). In addition, We consider two different $\sigma(\boldsymbol{X}_i)$ functions, i.e., the homogeneous case with $\sigma(\boldsymbol{X}_i) = 1$, and the heterogenous case with $\sigma(\boldsymbol{X}_i) = 0.5 + (X_{i1} - X_{i2})^2$. The results of Model I and II are given in Table 4.1 and 4.2 respectively.

Table 4.1: Summary result of Model I with constant propensity scores. LS stands for lsA-learning. P(0.5) and P(0.25) stand for robust regression with pinball loss and parameter $\tau = 0.5$ and 0.25, respectively. Huber stands for robust regression with Huber loss, where parameter α is tuned automatically with R function rlm. $\delta_{0.5}$ is multiplied by 10.

		Homogeneous Error								
		Normal			**Log-Normal**			**Cauchy**		
n	method	mse	PCD	$\delta_{0.5}$	mse	PCD	$\delta_{0.5}$	mse	PCD	$\delta_{0.5}$
100	LS	1.32 (0.040)	80.7	1.06	2.36 (0.081)	75.7	1.57		58.4	3.75
	P(0.5)	1.44 (0.042)	80.1	1.13	1.73 (0.051)	78.0	1.31	2.69 (0.077)	75.2	1.63
	P(0.25)	1.90 (0.057)	78.3	1.34	1.63 (0.051)	79.0	1.29	5.29 (0.168)	70.4	2.25
	Huber	1.15 (0.034)	81.9	0.93	1.45 (0.044)	79.9	1.13	2.61 (0.072)	74.9	1.66
200	LS	0.68 (0.021)	85.6	0.59	1.10 (0.033)	82.0	0.91		58.7	3.70
	P(0.5)	0.73 (0.021)	85.3	0.62	0.78 (0.021)	84.1	0.70	1.23 (0.037)	81.3	0.99
	P(0.25)	0.92 (0.028)	84.0	0.75	0.70 (0.023)	86.0	0.59	2.48 (0.079)	75.7	1.64
	Huber	0.58 (0.017)	86.8	0.50	0.66 (0.018)	85.5	0.58	1.24 (0.035)	80.8	1.03
400	LS	0.33 (0.009)	90.3	0.26	0.56 (0.016)	87.1	0.46		59.2	3.61
	P(0.5)	0.35 (0.010)	90.0	0.29	0.37 (0.010)	89.0	0.34	0.56 (0.016)	87.1	0.48
	P(0.25)	0.43 (0.013)	89.1	0.34	0.33 (0.010)	90.7	0.25	1.16 (0.037)	82.9	0.86
	Huber	0.28 (0.008)	91.1	0.22	0.31 (0.009)	90.2	0.27	0.58 (0.017)	86.7	0.49
800	LS	0.17 (0.005)	93.2	0.13	0.26 (0.008)	90.9	0.23		59.4	3.59
	P(0.5)	0.17 (0.005)	93.1	0.13	0.19 (0.005)	92.1	0.17	0.29 (0.009)	90.7	0.24
	P(0.25)	0.22 (0.007)	92.4	0.16	0.18 (0.006)	93.6	0.12	0.59 (0.019)	87.3	0.48
	Huber	0.14 (0.004)	93.8	0.11	0.16 (0.005)	93.1	0.14	0.29 (0.008)	90.5	0.25
		Heterogenous Error								
		Normal			**Log-Normal**			**Cauchy**		
n	method	mse	PCD	$\delta_{0.5}$	mse	PCD	$\delta_{0.5}$	mse	PCD	$\delta_{0.5}$
100	LS	3.24 (0.110)	74.7	1.70	8.98 (0.561)	68.6	2.44		56.2	4.05
	P(0.5)	1.70 (0.060)	80.5	1.08	1.80 (0.064)	80.1	1.08	3.45 (0.124)	75.1	1.69
	P(0.25)	2.50 (0.085)	77.4	1.42	2.51 (0.079)	76.8	1.46	9.13 (0.341)	67.2	2.66
	Huber	1.70 (0.057)	80.4	1.10	1.87 (0.063)	79.2	1.16	4.27 (0.155)	72.8	1.93
200	LS	1.54 (0.050)	80.6	1.06	4.71 (0.244)	73.4	1.85		55.2	4.17
	P(0.5)	0.78 (0.028)	86.7	0.53	0.90 (0.032)	85.3	0.63	1.49 (0.052)	81.9	0.95
	P(0.25)	1.16 (0.039)	83.5	0.81	1.23 (0.039)	82.0	0.91	3.95 (0.150)	73.2	1.90
	Huber	0.77 (0.025)	86.4	0.55	0.94 (0.032)	84.5	0.69	1.94 (0.071)	79.3	1.19
400	LS	0.80 (0.026)	86.0	0.58	2.69 (0.136)	77.8	1.34		54.7	4.26
	P(0.5)	0.39 (0.013)	90.5	0.27	0.44 (0.017)	89.6	0.32	0.71 (0.024)	86.9	0.50
	P(0.25)	0.56 (0.019)	88.8	0.37	0.66 (0.020)	86.9	0.50	1.70 (0.055)	79.6	1.17
	Huber	0.38 (0.012)	90.4	0.27	0.48 (0.017)	88.8	0.36	0.91 (0.029)	84.9	0.65
800	LS	0.41 (0.013)	89.9	0.29	1.35 (0.150)	83.1	0.82		56.5	4.00
	P(0.5)	0.18 (0.006)	93.6	0.12	0.20 (0.007)	92.6	0.16	0.36 (0.013)	91.0	0.25
	P(0.25)	0.28 (0.009)	92.2	0.18	0.31 (0.010)	90.8	0.24	0.89 (0.031)	85.8	0.60
	Huber	0.19 (0.006)	93.3	0.13	0.22 (0.007)	92.1	0.18	0.47 (0.017)	89.2	0.34

Table 4.2: Summary result of Model II with constant propensity scores. LS stands for lsA-learning. P(0.5) and P(0.25) stand for robust regression with pinball loss and parameter $\tau = 0.5$ and 0.25, respectively. Huber stands for robust regression with Huber loss, where parameter α is tuned automatically with R function rlm. $\delta_{0.5}$ is multiplied by 10.

		Homogeneous Error								
		Normal			**Log-Normal**			**Cauchy**		
n	method	mse	PCD	$\delta_{0.5}$	mse	PCD	$\delta_{0.5}$	mse	PCD	$\delta_{0.5}$
100	LS	0.24 (0.006)	91.1	0.21	1.23 (0.061)	82.4	0.87		58.6	3.73
	P(0.5)	0.36 (0.010)	89.0	0.32	0.39 (0.012)	88.8	0.34	0.80 (0.024)	84.2	0.69
	P(0.25)	0.45 (0.012)	87.8	0.40	0.13 (0.004)	93.4	0.12	2.37 (0.083)	76.0	1.49
	Huber	0.25 (0.007)	90.8	0.22	0.31 (0.010)	90.3	0.26	0.99 (0.029)	82.4	0.84
200	LS	0.11 (0.003)	93.7	0.10	0.52 (0.018)	87.3	0.45		58.7	3.69
	P(0.5)	0.17 (0.005)	92.4	0.16	0.17 (0.005)	92.4	0.15	0.32 (0.009)	89.5	0.30
	P(0.25)	0.20 (0.005)	91.8	0.18	0.06 (0.002)	95.6	0.05	1.03 (0.033)	82.1	0.88
	Huber	0.12 (0.003)	93.6	0.11	0.13 (0.003)	93.5	0.12	0.43 (0.013)	87.9	0.40
400	LS	0.05 (0.001)	95.7	0.05	0.26 (0.008)	90.7	0.23		59.4	3.60
	P(0.5)	0.09 (0.002)	94.5	0.08	0.09 (0.002)	94.5	0.08	0.15 (0.004)	92.8	0.14
	P(0.25)	0.10 (0.002)	94.2	0.09	0.03 (0.001)	96.9	0.02	0.44 (0.012)	87.9	0.39
	Huber	0.06 (0.001)	95.5	0.05	0.06 (0.002)	95.4	0.06	0.21 (0.006)	91.6	0.19
800	LS	0.03 (0.001)	96.9	0.03	0.13 (0.004)	93.5	0.11		59.4	3.58
	P(0.5)	0.04 (0.001)	96.1	0.04	0.04 (0.001)	96.2	0.04	0.07 (0.002)	95.1	0.06
	P(0.25)	0.05 (0.001)	95.8	0.05	0.01 (0.000)	97.9	0.01	0.20 (0.005)	91.5	0.19
	Huber	0.03 (0.001)	96.8	0.03	0.03 (0.001)	96.8	0.03	0.10 (0.002)	94.2	0.09
		Heterogenous Error								
		Normal			**Log-Normal**			**Cauchy**		
n	method	mse	PCD	$\delta_{0.5}$	mse	PCD	$\delta_{0.5}$	mse	PCD	$\delta_{0.5}$
100	LS	1.97 (0.072)	79.8	1.13	7.75 (0.514)	70.4	2.22		56.4	4.02
	P(0.5)	0.84 (0.029)	86.1	0.55	1.21 (0.045)	84.3	0.74	1.82 (0.071)	80.5	1.07
	P(0.25)	1.37 (0.049)	82.1	0.90	1.56 (0.051)	80.5	1.04	6.20 (0.261)	69.8	2.25
	Huber	0.84 (0.031)	85.9	0.57	1.33 (0.046)	82.8	0.85	2.69 (0.106)	77.0	1.42
200	LS	0.99 (0.035)	84.7	0.66	4.16 (0.237)	75.2	1.62		55.1	4.19
	P(0.5)	0.41 (0.014)	90.2	0.28	0.58 (0.024)	89.4	0.37	0.79 (0.030)	86.7	0.52
	P(0.25)	0.64 (0.021)	87.4	0.45	0.74 (0.024)	86.1	0.54	2.48 (0.096)	76.9	1.40
	Huber	0.39 (0.013)	90.3	0.27	0.69 (0.027)	87.7	0.45	1.17 (0.044)	83.4	0.78
400	LS	0.51 (0.018)	89.0	0.35	2.48 (0.133)	79.3	1.20		54.7	4.25
	P(0.5)	0.20 (0.007)	93.2	0.14	0.29 (0.011)	92.6	0.17	0.32 (0.011)	91.2	0.22
	P(0.25)	0.30 (0.009)	91.3	0.22	0.39 (0.012)	89.9	0.28	0.99 (0.030)	83.0	0.78
	Huber	0.20 (0.007)	93.2	0.14	0.34 (0.012)	91.4	0.22	0.53 (0.016)	88.4	0.37
800	LS	0.25 (0.008)	92.2	0.17	1.25 (0.159)	84.2	0.73		56.4	4.00
	P(0.5)	0.10 (0.004)	95.3	0.07	0.14 (0.006)	94.7	0.09	0.16 (0.006)	93.9	0.11
	P(0.25)	0.14 (0.005)	94.0	0.10	0.18 (0.006)	92.9	0.14	0.49 (0.015)	88.0	0.39
	Huber	0.09 (0.004)	95.3	0.06	0.17 (0.006)	93.9	0.11	0.26 (0.009)	91.8	0.19

Comparison is made among four methods. They are: lsA-learning, robust regression with $\rho_{0.5}$ (RR($\rho_{0.5}$)), robust regression with $\rho_{0.25}$ (RR($\rho_{0.25}$)), and robust regression with Huber loss (RR(H)). The error terms ϵ_i are taken as standard i.i.d. normal, log-normal or Cauchy distribution, and independent with both A and $\boldsymbol{X}$. It is easy to check that the conditional independence assumption $\epsilon \perp A|\boldsymbol{X}$ is satisfied, and $g_\mu^{\text{opt}} = g_\tau^{\text{opt}} = 1\{\boldsymbol{\beta}_0^\intercal \tilde{\boldsymbol{X}}_i > 0\}$. We consider four different sample sizes 100, 200, 400 and 800. To evaluate the performance of each method, we compare three groups of criteria: (1) the mean squared error $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2^2$ (mse), which measures the distance between estimated parameters and the true parameter $\boldsymbol{\beta}_0$; (2) the percentage of making correct decisions (PCD), which are calculated based on a validation set with 10000 observations. Specifically, we take the formula $100 * \left(1 - \sum_{i=1}^{N_T} |1\{\hat{\boldsymbol{\beta}}^\intercal \tilde{\boldsymbol{X}}_i > 0\} - 1\{\boldsymbol{\beta}_0^\intercal \tilde{\boldsymbol{X}}_i > 0\}|/N_T\right)$ with $N_T = 10000$; (3) the differences of $V_\mu(g)$ and $V_{0.5-q}(g)$ between the optimal ITR and the estimated ITR, where $\delta_\mu = V_\mu(g_\mu^{\text{opt}}) - V_\mu(\hat{g})$ and $\delta_\tau = V_{\tau-q}(g_\mu^{\text{opt}}) - V_{\tau-q}(\hat{g})$, $\forall \tau \in (0, 1)$. $V_\mu(g)$ and $V_{\tau-q}(g)$ (defined in Section 4.1) are estimated from the validation set as well, and they evaluate the overall performance of an ITR g, where the former one focuses on the response's mean and the latter one focuses on the response's τ-th quantile. Under our setting, $\delta_\mu = \delta_{0.5}$ when they both exists. Thus, only $\delta_{0.5}$ is reported. For each scenario, we take 1000 replications. All numbers in the tables are based on the sample average of all replications. We further report the standard errors of mse to evaluate the variability of the corresponding statistics.

The simulation results under constant and non-constant propensity scores are similar. Thus, for brevity, we only report the constant case and allocate the result of non-constant case to the Supplement. When the propensity score is constant, lsA-learning is equivalent to both Q- and A-learning under our setting. If we compare the performance of the methods under homogeneous and heterogeneous errors, the first thing we find is that lsA-learning works much worse under the heterogeneous errors, while all other methods are generally less affected by the heterogeneity of the errors. When the baseline function is misspecified as in Model I, under the homogeneous normal errors, RR(H) works slightly better than lsA-learning, while RR($\rho_{0.25}$) works the worst. However, the difference in general is small. For the homogeneous log-normal errors, again RR(H) works the best, while RR($\rho_{0.5}$) and RR($\rho_{0.25}$) have similar performance, and lsA-learning works the worst. Under the homogeneous Cauchy errors, the lsA-learning is no longer consistent, and its mse explodes. The actual numbers are too large and thus leave as blank in Table 4.1 and 4.2. Furthermore, with the Cauchy errors, the PCD of lsA-learning are less than 60% under all scenarios, while other methods' PCD can be as high as 90%. When baseline function is correctly specified as in Model II, under homogeneous normal errors, lsA-learning performs the best. However, in this case RR(H) also has a very close performance, and thus makes no difference from a practical point of view to choose between these two methods. The results of Model II under other cases draw similar conclusion as Model I. To sum up, the overall conclusion is that, under the conditional independence assumption, the proposed robust regression method RR(M) is more efficient than Q-, A- and lsA-learning in the circumstances when observations have skewed, heterogeneous or heavy-tailed errors. On the other hand, when the error terms indeed follows i.i.d. normal distribution, the loss of efficiency of RR(M) is not significant. This is especially true when Huber loss is applied.

4.4.2 Simulation Study II: error terms interactive with treatment

We consider the following model with p=2,

$$Y_i = 1 + 0.5\sin[\pi(X_{i1} - X_{i2})] + 0.25(1 + X_{i1} + 2X_{i2})^2 + (A_i - \pi(\boldsymbol{X}_i))\boldsymbol{\theta}_0^\intercal \tilde{\boldsymbol{X}}_i + \sigma(\boldsymbol{X}_i, A_i)\epsilon_i,$$

where $\boldsymbol{X}_i = (X_{i1}, X_{i2})^\intercal$, $\tilde{\boldsymbol{X}}_i = (1, \boldsymbol{X}_i^\intercal)^\intercal$, $\sigma(\boldsymbol{X}_i, A_i) = 1 + A_i d_0 X_{i1}^2$, $\boldsymbol{\theta}_0^\intercal = (0.5, 2, -1)$ and X_{ik} are i.i.d. Uniform[-1,1].

Similar as Section 4.1, we take linear forms for both the baseline and the contrast functions, where $\varphi(\boldsymbol{X}; \boldsymbol{\gamma}) = \boldsymbol{\gamma}^\intercal \tilde{\boldsymbol{X}}$, $C(\boldsymbol{X}; \boldsymbol{\beta}) = \boldsymbol{\beta}^\intercal \boldsymbol{W}$ and $\boldsymbol{W} = (\tilde{\boldsymbol{X}}, X_1^2, X_2^2, X_1 X_2)$. The error terms ϵ_i follows i.i.d. N(0,1) or Gamma(1,1)-1 distribution. The propensity scores $\pi(\cdot)$ are known, and we consider both the constant case $\pi(\boldsymbol{X}_i) = 0.5$ and the non-constant case $\pi(\boldsymbol{X}_i) = \text{logit}(\boldsymbol{X}_{i1} - \boldsymbol{X}_{i2})$. We report only the result of the constant case (Table 4.3), and allocate the non-constant case to Supplement.

Table 4.3: Summary results with constant propensity scores when errors interacted with treatment. Least square stands for lsA-learning. Pinball(0.5) stands for robust regression with pinball loss and parameter $\tau = 0.5$. Pinball(0.25) stands for robust regression with pinball loss and parameter $\tau = 0.25$.

			Least Square					**Pinball(0.5)**					**Pinball(0.25)**				
Error	d_0	n	mse	PCD	δ_μ	$\delta_{0.5}$	$\delta_{0.25}$	mse	PCD	δ_μ	$\delta_{0.5}$	$\delta_{0.25}$	mse	PCD	δ_μ	$\delta_{0.5}$	$\delta_{0.25}$
Normal	0	100	2.86	89.1	0.04	0.04	0.72	4.72	86.0	0.07	0.07	0.74	5.96	84.7	0.08	0.08	0.75
		200	1.31	92.4	0.02	0.02	0.69	2.14	90.4	0.03	0.03	0.71	2.94	88.8	0.04	0.04	0.72
		400	0.64	94.7	0.01	0.01	0.68	1.06	93.2	0.02	0.02	0.69	1.41	92.2	0.02	0.02	0.70
		800	0.31	96.3	0.00	0.00	0.68	0.52	95.2	0.01	0.01	0.68	0.66	94.6	0.01	0.01	0.69
	5	100	13.92	78.5	0.16	0.16	1.27	10.41	78.5	0.16	0.16	1.23	15.95	76.3	0.25	0.25	1.12
		200	6.67	84.5	0.09	0.09	1.20	5.65	83.4	0.10	0.10	1.15	9.31	82.5	0.18	0.18	1.04
		400	3.34	89.3	0.05	0.05	1.14	3.55	87.2	0.06	0.06	1.09	6.41	86.0	0.15	0.15	1.01
		800	1.58	92.7	0.02	0.02	1.10	2.61	89.5	0.05	0.05	1.05	5.01	88.1	0.14	0.14	0.99
	10	100	37.94	70.3	0.27	0.27	1.73	15.76	74.3	0.21	0.21	1.61	39.89	71.0	0.38	0.38	1.20
		200	18.40	76.9	0.19	0.19	1.65	8.29	79.5	0.15	0.15	1.51	27.51	76.0	0.33	0.33	1.08
		400	9.11	82.3	0.12	0.12	1.60	5.40	83.7	0.10	0.10	1.40	23.61	78.8	0.30	0.30	1.03
		800	4.56	87.7	0.06	0.06	1.52	4.14	86.3	0.07	0.07	1.31	21.52	80.3	0.26	0.26	1.02
Gamma	0	100	2.76	89.2	0.04	0.04	0.04	3.92	87.5	0.05	0.05	0.05	3.76	88.1	0.05	0.05	0.05
		200	1.30	92.4	0.02	0.02	0.02	1.86	91.3	0.03	0.03	0.03	1.85	91.3	0.03	0.03	0.03
		400	0.63	94.6	0.01	0.01	0.01	0.97	93.5	0.01	0.01	0.01	0.94	93.6	0.02	0.02	0.02
		800	0.31	96.3	0.01	0.01	0.01	0.49	95.3	0.01	0.01	0.01	0.47	95.5	0.01	0.01	0.01
	5	100	13.35	79.9	0.14	0.17	0.29	8.10	81.8	0.14	0.11	0.16	8.76	82.4	0.21	0.11	0.09
		200	6.63	84.9	0.09	0.12	0.25	4.59	85.9	0.10	0.06	0.10	5.50	86.2	0.18	0.08	0.05
		400	3.27	89.0	0.05	0.08	0.21	3.04	88.4	0.08	0.03	0.07	4.24	88.3	0.16	0.06	0.04
		800	1.61	92.6	0.02	0.04	0.16	2.35	89.4	0.07	0.03	0.06	3.50	89.4	0.15	0.06	0.03
	10	100	38.37	70.4	0.28	0.36	0.92	14.62	78.3	0.23	0.17	0.55	28.46	76.9	0.41	0.14	0.26
		200	19.57	76.3	0.20	0.29	0.88	8.93	84.4	0.17	0.08	0.43	22.67	80.1	0.38	0.10	0.20
		400	9.33	82.9	0.11	0.22	0.83	6.77	87.8	0.13	0.04	0.38	20.85	81.2	0.35	0.08	0.19
		800	4.64	87.7	0.06	0.18	0.79	6.05	89.3	0.12	0.03	0.37	20.02	81.9	0.32	0.06	0.18

We compare the performance of three methods: lsA-learning, robust regression with $\rho_{0.5}$ (RR($\rho_{0.5}$)) and robust regression with $\rho_{0.25}$ (RR($\rho_{0.25}$)). We consider four different sample sizes 100, 200, 400 and 800. For each scenario, we again simulate 1000 replications. Similar as Section 4.1, we evaluate the performance of the method based on three criteria: the mean squared error $\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|_2^2$ (mse), the percentage of making correct decisions (PCD) and the value differences (δ_μ, $\delta_{0.5}$ and $\delta_{0.25}$). However, when the error is dependent with A, the true parameter $\boldsymbol{\beta}_0$ associated with each method is different, which correspond to the true contrast function of conditional mean, conditional median and conditional 25% quantile, respectively. Thus, $\boldsymbol{\beta}_0 = (\boldsymbol{\theta}_0^\intercal, 0, 0, 0)^\intercal$ for lsA-learning, $\boldsymbol{\beta}_0 = (\boldsymbol{\theta}_0^\intercal, d_0 F_\epsilon^{-1}(0.5), 0, 0)^\intercal$ for RR($\rho_{0.5}$) and $\boldsymbol{\beta}_0 = (\boldsymbol{\theta}_0^\intercal, d_0 F_\epsilon^{-1}(0.25), 0, 0)^\intercal$ for RR($\rho_{0.25}$).

We first take a quick look at the mse columns of Table 4.3. When $d_0 = 0$, the conditional independence assumption $\epsilon \perp A|\boldsymbol{X}$ is satisfied. Thus, lsA-learning, RR($\rho_{0.5}$) and RR($\rho_{0.25}$) are all consistent. Their corresponding mean square errors converge to zero when sample size goes to infinity. On the other hand, when $d_0 \neq 0$, only lsA-learning is consistent. The mses of RR($\rho_{0.5}$) and RR($\rho_{0.25}$) converge to positive constants due to the underlying biases. However, in this case when sample size is small, the mse of lsA-learning can be larger than the other two methods. This is due to the fact that lsA-learning is inefficient under the heteroscedastic or skewed errors. Another observation we have is overall lsA-learning, RR($\rho_{0.5}$) and RR($\rho_{0.25}$) perform best at the columns δ_μ, $\delta_{0.5}$ and $\delta_{0.25}$ accordingly. The reason is given in the Remark under Theorem 4.2, which shows that $\hat{g}^R_{\rho(\tau)}$ ($\triangleq 1\{C(\boldsymbol{x}; \hat{\boldsymbol{\beta}}^R_{\rho(\tau)}) > 0\}$) in general approximates the unknown optimal ITR g_τ^{opt} even when the conditional independence assumption $\epsilon \perp A|\boldsymbol{X}$ does not hold.

4.5 Application to AIDS study

We illustrate the proposed robust regression method to data from AIDS Clinical Trials Group Protocol 175 (ACTG175), which has been previously studied in Leon et al. (2003); Tsiatis et al. (2008); Zhang et al. (2008); Lu et al. (2013). In the study, 2139 HIV-infected subjects were randomized to four different treatment groups in equal proportions, and the treatment groups are zidovudine (ZDV) monotherapy, ZDV + didanosine (ddI), ZDV + zalcitabine, and ddI monotherapy (Tsiatis et al., 2008). Following Lu et al. (2013), we choose CD4 count (cells/mm^3) at 20$\pm$5 weeks post-baseline as the primary continuous outcome Y, and include five continuous covariates and seven binary covariates as our covariates. They are: 1. age (years), 2. weight (kg), 3. karnof=Karnofsky score (scale of 0-100), 4. cd40=CD4 count (cells/mm^3)

at baseline, 5. cd80=CD8 count (cells/mm^3) at baseline, 6. hemophilia=hemophilia (0=no, 1=yes), 7. homosexuality=homosexual activity (0=no, 1=yes), 8. drugs=history of intravenous drug use (0=no, 1=yes), 9. race (0=white, 1=non-white), 10. gender (0=female, 1=male), 11. str2= antiretroviral history (0=naive, 1=experienced), and 12. sympton=symptomatic status (0=asymptomatic, 1=symptomatic). For brevity, we only compare the treatment ZDV + didanosine (ddI) ($A = 1$) and ZDV + zalcitabine ($A = 0$), and restrict our samples to subjects receiving these two treatments. Thus, the propensity scores $\pi(\boldsymbol{X}_i) \equiv 0.5$ in our restricted samples as the patients are assigned into one of two treatments with equal probability.

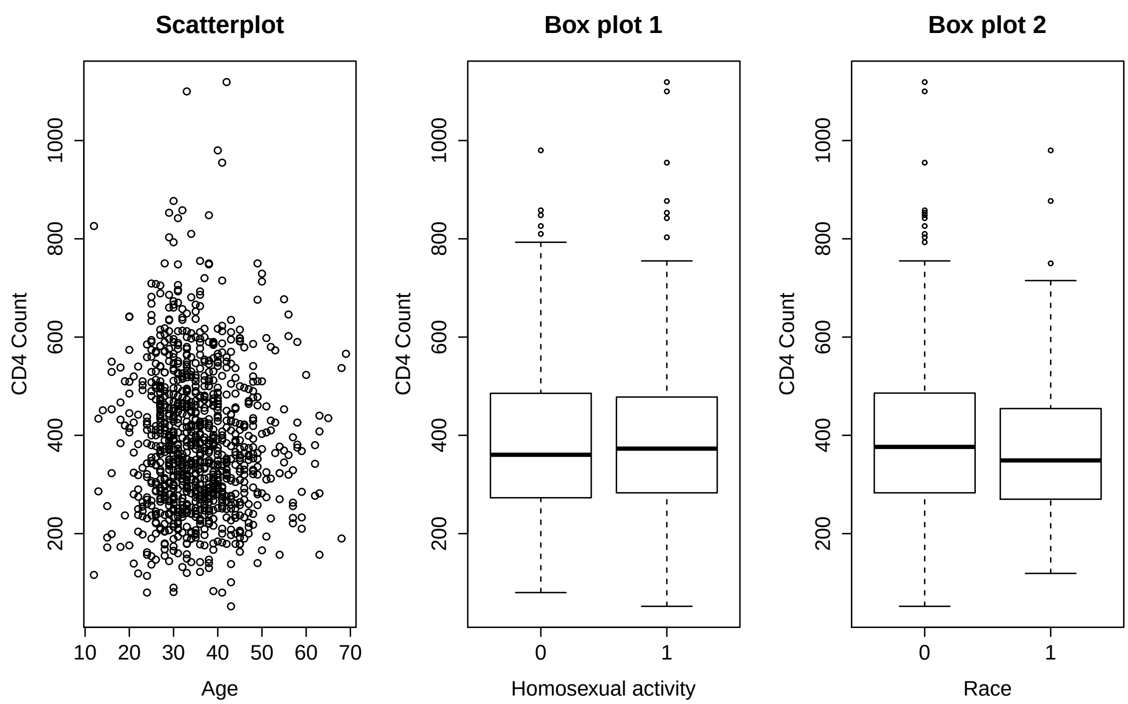


Figure 4.3: Scatterplot and box plots of response Y (CD4 count (cells/mm^3) at 20 ± 5 weeks). Left Panel: Scatterplot of Y and covariate age; Middle Panel: Box plot of Y stratified by covariate homosexuality (homosexual activity). Right Panel: Box plot of Y stratified by covariate race.

In our analysis, we assume linear models for both the baseline and the contrast functions.

For interpretability, we keep the response Y (the CD4 count) at its original scale, which is also consistent with the way clinicians think about the outcome in practice (Tsiatis et al., 2008). In Figure 4.3, we plot the scatterplot of response Y against age, and also its box plots stratified by class variable homosexuality (homosexual activity) and race. From Figure 4.3, clearly Y exhibits a skewed and heterogeneous distribution. With some preliminary experiments (fitting full model with lsA-learning and RR(M)), we find that only covariates age, homosexuality and race may possibly interact with the treatment. So in our final model, only these three covariates are included in the contrast function, while at the same time we still keep all twelve covariates in the baseline function. The estimated coefficients associated with their corresponding standard errors and p-values are given in Table 4.4, where standard errors are estimated with 1000 bootstrap samples (parametric bootstrap) and p-values are calculated with normal approximation. Only coefficients included in the contrast function are shown.

Table 4.4: Analysis results for AIDS data. Est. stands for estimate; SE stands for standard error; PV stands for p-value. All p-values which are significant at level 0.05 are highlighted.

	Least Square			**Pinball(0.5)**			**Pinball(0.25)**			**Huber**		
Variable	Est.	SE	PV	Est.	SE	PV	Est.	SE	PV	Est.	SE	PV
intercept	-42.61	33.25	0.200	-33.45	38.49	0.385	-35.77	38.72	0.356	-42.76	32.08	0.183
age	3.13	0.84	**0.000**	2.62	0.98	**0.008**	2.46	1.04	**0.018**	2.80	0.79	**0.000**
homosexuality	-40.66	17.11	**0.018**	-33.18	17.56	0.059	-35.38	18.26	0.053	-27.33	15.49	0.078
race	-25.70	16.91	0.129	-33.56	17.54	0.056	-34.21	17.94	0.057	-25.29	15.55	0.104

From Tables 4.4, we make the following observations. First, lsA-learning (equivalent to Q- and A-learning with this model setting) and robust regression with pinball loss and Huber loss all have estimates with the exact same signs. Second, the estimated coefficients are distinguishable across different methods. Third, the covairiate homosexuality is significant under lsA-learning, but it is not significant under robust regression with either pinball losses or Huber loss, when the significant level α is set to 0.05. Fourth, the covariate age is significant under all methods.

We could further estimate the values ($V_\mu(g)$) associated with each method by the inverse probability of treatment weighted estimator (IPTW) (Robins, 2000), where

$$\hat{V}_\mu(\hat{g}) = \frac{\sum_{i=1}^n 1\{A_i = \hat{g}(\boldsymbol{X}_i)\}Y_i/p(A_i|\boldsymbol{X}_i)}{\sum_{i=1}^n 1\{A_i = \hat{g}(\boldsymbol{X}_i)\}/p(A_i|\boldsymbol{X}_i)}.$$

and $p(A_i|\boldsymbol{X}_i) \equiv 0.5$. The estimates are calculated based on 10-fold CV, where we randomly partition the whole data set into 10 approximately equal size subsamples. We apply stratified random sampling to guarantee that each subsample has approximately equal numbers of subjects receiving each treatment. Thus, the condition $p(A_i|\boldsymbol{X}_i) \equiv 0.5$ is always satisfied. We repeat the above process 1000 times to generate the histograms of $\hat{V}_\mu(\hat{g})$ (Figure 4.4).

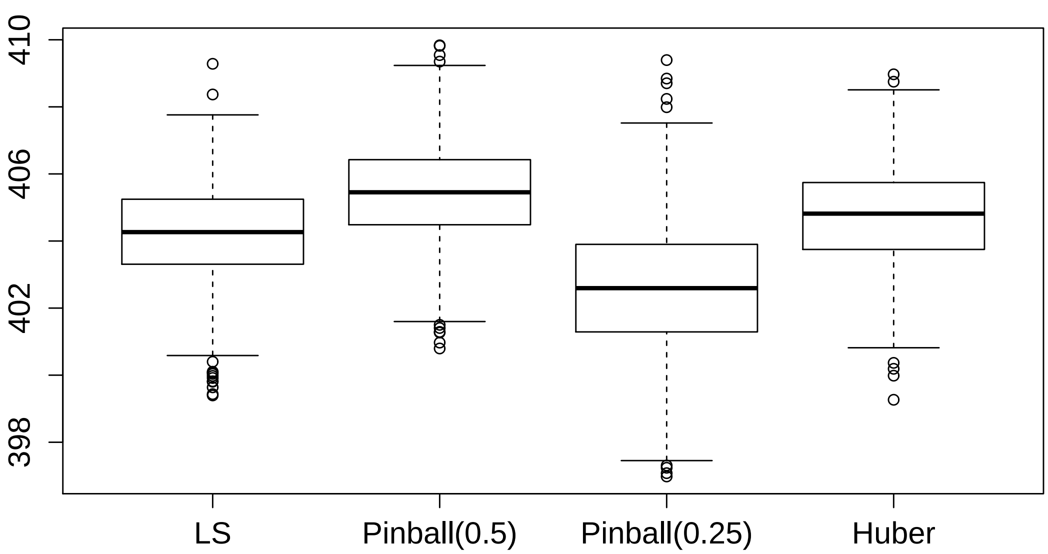


Figure 4.4: Distributions of estimated values (IPWE) of different methods based on 10-fold cross validation.

From Figure 4.4, robust regression with $\rho_{0.5}$ and Huber loss perform slightly better than lsA-learning, while robust regression with $\rho_{0.25}$ performs the worst. We conduct KCI-test to check the conditional independence assumption $\epsilon \perp A|\boldsymbol{X}$, and the result is given in Table 4.5. For the estimates of RR($\rho(0.5)$) and RR(H), the conditional independence assumption holds at the significance level of 0.05, so the estimated ITR can be thought to maximize $V_\mu(g)$. On the other hand, this assumption doesn't hold for RR($\rho(0.25)$), and its estimated ITR doesn't maximize $V_\mu(g)$, instead it approximately maximizes $V_{0.25-q}(g)$. This partly explains why RR($\rho_{0.25}$) performs the worst in Figure 4.4. Again, as RR($\rho(0.5)$) and RR(H) are more robust against heterogeneous, right skewed errors comparing with the least square method, they slightly outperform lsA-learning in term of $V_\mu(g)$.

Table 4.5: Result of hypothesis test for conditional independence assumption $\epsilon \perp A|\boldsymbol{X}$.

	Pinball(0.5)		**Pinball(0.25)**		**Huber**	
	Stat.	P-Value	Stat.	P-Value	Stat.	P-Value
KCI-test	36.319	0.060	67.249	**0.002**	33.214	0.083

4.6 Discussion

In this chapter, we propose a new general loss based robust regression framework for estimating the optimal individualized treatment rules. This new method has the desired property to be robust against the skewed, heterogeneous, heavy-tailed errors and outliers. And similar as A-learning, it produces consistent estimates of the optimal ITR even when the baseline function is misspecified. However, the consistency of the proposed method does require the key conditional independence assumption $\epsilon \perp A|\boldsymbol{X}$, which is somewhat stronger than the condition needed for the consistency of Q- and A-learning ($\mathrm{E}(\epsilon|\boldsymbol{X}, A) = 0$). So there are situations when the classical Q- and A-learning are more appropriate to apply. Furthermore, we also point out in the article that when pinball loss ρ_τ is chosen and the assumption $\epsilon \perp A|\boldsymbol{X}$ doesn't hold, the estimated ITR approximately maximize the conditional τ-th quantile and thus maximize $V_{\tau-q}(g)$. From a practice point of view, there are situations when maximizing $V_{\tau-q}(g)$ is a much more reasonable approach comparing with maximizing $V_\mu(g)$, especially when the conditional distribution of response Y is highly skewed to one side.

In practice, there are cases when multiple treatment groups need to be compared simultaneously. For brevity, we have limited our discussion to two treatment groups. However, the proposed method can be readily extended to multiple cases by just replacing equation (4.8) with the following more complex form,

$$L_{3n}(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \frac{1}{n}\sum_{i=1}^{n} M\left[Y_i - \varphi(\boldsymbol{X}_i; \boldsymbol{\gamma}) - \sum_{k=1}^{K-1}(I(A_i = k) - \pi_k(\boldsymbol{X}_i))C_k(\boldsymbol{X}_i; \boldsymbol{\beta}_k)\right],$$

where $\mathcal{A} = \{1, \ldots, K\}$, K-th treatment is the baseline treatment, $\pi_k(\boldsymbol{X}_i) = \Pr(A_i = k|\boldsymbol{X}_i)$ and $C_k(\boldsymbol{X}_i; \boldsymbol{\beta}_k)$ denotes the contrast function comparing k-th treatment and the baseline treatment. All Theorems can be easily extended to this multiple treatments setting as well.

When the dimension of prognostic variables is high, regularized regression is needed in order to produce parsimonious yet interpretable individualized treatment rules. Essentially this is a variable selection problem in the context of M-estimator, which has been previously studied in Li et al. (2011); Wu and Liu (2009), etc. This is an interesting topic that needs further investigation. Another interesting direction is to extend the current method to the multi-stage setting, where sequential decisions are made along the time line.

4.7 Supplement

4.7.1 Proof of Asymptotic Properties of Robust Regression

Define $\Delta C(\boldsymbol{x}; \boldsymbol{\beta}) = C(\boldsymbol{x}; \boldsymbol{\beta}) - C(\boldsymbol{x}; \boldsymbol{\beta}_0)$. Assume $\boldsymbol{\gamma} \in \Gamma$, $\boldsymbol{\beta} \in \mathcal{B}$ and $\boldsymbol{\gamma}'$ is any arbitrary fix point in Γ.

Regularity conditions A:

(A1) $\{(Y_i, \boldsymbol{X}_i, A_i, \epsilon_i), i = 1, \ldots, n\}$ are i.i.d random variables.

(A2) $\epsilon_i \perp A_i|\boldsymbol{X}_i \ \forall i = 1, \ldots, n$.

(A3) $\mathrm{E}|\Delta C(\boldsymbol{X}_i;\boldsymbol{\beta})| < \infty \ \forall \boldsymbol{\beta} \in \mathcal{B}$.

(A4) $\Pr\{\boldsymbol{x} \in \mathcal{X} : \ \Delta C(\boldsymbol{x};\boldsymbol{\beta}) \neq 0\} > 0$ for all $\boldsymbol{\beta} \neq \boldsymbol{\beta}_0$.

(A5) $\mathrm{E}|\varphi(\boldsymbol{X}_i;\boldsymbol{\gamma})| < \infty \ \forall \boldsymbol{\gamma} \in \Gamma$.

(A6) $G_2(\boldsymbol{\gamma})$ has unique minimizer $\boldsymbol{\gamma}^*$, where $G_2(\boldsymbol{\gamma})$ is the pointwise limit of $L_{3n}(\boldsymbol{\beta}_0,\boldsymbol{\gamma}) - L_{3n}(\boldsymbol{\beta}_0,\boldsymbol{\gamma}')$ in probability.

(A7) $L_{3n}(\boldsymbol{\beta},\boldsymbol{\gamma})$ is strictly convex with respect to $(\boldsymbol{\beta},\boldsymbol{\gamma})$.

(A8) $\epsilon|\boldsymbol{X} = \boldsymbol{x}$ has nonzero density on $\mathbb{R}$ for almost all $\boldsymbol{x} \in \mathcal{X}$.

Lemma 4.1. $|\rho_\tau(x-y) - \rho_\tau(x)| \leq |y|$, *for all* $\tau \in (0,1)$.

Proof of Lemma 4.1.

$$\begin{aligned}|\rho_\tau(x-y) - \rho_\tau(x)| &= |\tau\left\{(x-y)_+ - x_+\right\} + (1-\tau)\left\{(x-y)_- - x_-\right\}| \\ &\leq |(x-y)_+ - x_+| + |(x-y)_- - x_-| = |y|\end{aligned}$$

□

Lemma 4.2.

$$\begin{aligned}\rho_\tau(x-y) - \rho_\tau(x) = &-\tau y 1\{x \geq 0\} + (1-\tau) y 1\{x < 0\} + (y-x)1\{x \geq 0\}1\{y > x\} \\ &+ (x-y)1\{x < 0\}1\{y < x\},\end{aligned}$$

for all $\tau \in (0,1)$.

Proof of Lemma 4.2. Denote $D = \rho_\tau(x-y) - \rho_\tau(x)$.

1. If $x \geq 0$, $y \leq 0 \Rightarrow D = -\tau y$;
2. If $x \geq 0$, $y > 0$, $|x| \geq |y| \Rightarrow D = -\tau y$;

3. If $x \geq 0$, $y > 0$, $|x| < |y| \Rightarrow D = -\tau y + (y - x)$;

4. If $x < 0$, $y \geq 0 \Rightarrow D = (1-\tau)y$;

5. If $x < 0$, $y < 0$, $|x| \geq |y| \Rightarrow D = (1-\tau)y$;

6. If $x < 0$, $y < 0$, $|x| < |y| \Rightarrow D = (1-\tau)y + (x - y)$;

Combining the above 6 cases, Lemma 4.2 is proved. $\square$

Proof of Theorem 4.1. Recall that the loss function defined in (4.8) takes the form

$$L_{3n}(\boldsymbol{\beta},\boldsymbol{\gamma}) = \frac{1}{n}\sum_{i=1}^{n}\rho_\tau\left[\varphi_0(\boldsymbol{X}_i) - \varphi(\boldsymbol{X}_i;\boldsymbol{\gamma}) + \epsilon_i - (A_i - \pi(\boldsymbol{X}_i))\Delta C(\boldsymbol{X}_i;\boldsymbol{\beta})\right].$$

By definition,

$$\begin{aligned}(\hat{\boldsymbol{\beta}}^R_{\rho(\tau)}, \hat{\boldsymbol{\gamma}}^R_{\rho(\tau)}) =& \text{argmin}_{(\boldsymbol{\beta},\boldsymbol{\gamma})} L_{3n}(\boldsymbol{\beta},\boldsymbol{\gamma}) - L_{3n}(\boldsymbol{\beta}_0,\boldsymbol{\gamma}') \\ =& \text{argmin}_{(\boldsymbol{\beta},\boldsymbol{\gamma})}\left[L_{3n}(\boldsymbol{\beta},\boldsymbol{\gamma}) - L_{3n}(\boldsymbol{\beta}_0,\boldsymbol{\gamma})\right] + \left[L_{3n}(\boldsymbol{\beta}_0,\boldsymbol{\gamma}) - L_{3n}(\boldsymbol{\beta}_0,\boldsymbol{\gamma}')\right],\end{aligned}$$

Define

$$\begin{aligned}S_{1n}(\boldsymbol{\beta},\boldsymbol{\gamma}) =& L_{3n}(\boldsymbol{\beta},\boldsymbol{\gamma}) - L_{3n}(\boldsymbol{\beta}_0,\boldsymbol{\gamma}) = 1/n\sum_{i=1}^{n} d_{1i}; \\ S_{2n}(\boldsymbol{\beta},\boldsymbol{\gamma}) =& L_{3n}(\boldsymbol{\beta}_0,\boldsymbol{\gamma}) - L_{3n}(\boldsymbol{\beta}_0,\boldsymbol{\gamma}') = 1/n\sum_{i=1}^{n} d_{2i}\end{aligned}$$

where

$$\begin{aligned}d_{1i} =& \rho_\tau\left[\varphi_0(\boldsymbol{X}_i) - \varphi(\boldsymbol{X}_i;\boldsymbol{\gamma}) + \epsilon_i - (A_i - \pi(\boldsymbol{X}_i))\Delta C(\boldsymbol{X}_i;\boldsymbol{\beta})\right] - \rho_\tau\left[\varphi_0(\boldsymbol{X}_i) - \varphi(\boldsymbol{X}_i;\boldsymbol{\gamma}) + \epsilon_i\right], \\ d_{2i} =& \rho_\tau\left[\varphi_0(\boldsymbol{X}_i) - \varphi(\boldsymbol{X}_i;\boldsymbol{\gamma}) + \epsilon_i\right] - \rho_\tau\left[\varphi_0(\boldsymbol{X}_i) - \varphi(\boldsymbol{X}_i;\boldsymbol{\gamma}') + \epsilon_i\right].\end{aligned}$$

By Lemma 4.1, A3 and A5, $\mathrm{E}|d_{1i}| \leq \mathrm{E}|(A_i - \pi(\boldsymbol{X}_i))\Delta C(\boldsymbol{X}_i;\boldsymbol{\beta})| \leq \mathrm{E}|\Delta C(\boldsymbol{X}_i;\boldsymbol{\beta})| < \infty$ and $\mathrm{E}|d_{2i}| \leq \mathrm{E}|\varphi(\boldsymbol{X}_i;\boldsymbol{\gamma}) - \varphi(\boldsymbol{X}_i;\boldsymbol{\gamma}')| \leq \mathrm{E}|\varphi(\boldsymbol{X}_i;\boldsymbol{\gamma})| + \mathrm{E}|\varphi(\boldsymbol{X}_i;\boldsymbol{\gamma}')| < \infty$. Then, by Law of Large

Number, $\forall\, \boldsymbol{\beta} \in \mathcal{B}$, $\boldsymbol{\gamma} \in \Gamma$, we have $S_{1n}(\boldsymbol{\beta},\boldsymbol{\gamma}) \xrightarrow{p} G_1(\boldsymbol{\beta},\boldsymbol{\gamma}) \triangleq \mathrm{E}(D)$, and $S_{2n}(\boldsymbol{\beta},\boldsymbol{\gamma}) \xrightarrow{p} G_2(\boldsymbol{\gamma})$, where

$$D = \rho_\tau \left[\varphi_0(\boldsymbol{X}) - \varphi(\boldsymbol{X};\boldsymbol{\gamma}) + \epsilon - (A - \pi(\boldsymbol{X}))\Delta C(\boldsymbol{X};\boldsymbol{\beta})\right] - \rho_\tau \left[\varphi_0(\boldsymbol{X}) - \varphi(\boldsymbol{X};\boldsymbol{\gamma}) + \epsilon\right].$$

Below we show that a) $(\boldsymbol{\beta}_0,\boldsymbol{\gamma}^*)$ is the minimizer of $G_1(\boldsymbol{\beta},\boldsymbol{\gamma}) + G_2(\boldsymbol{\gamma})$, b) $(\boldsymbol{\beta}_0,\boldsymbol{\gamma}^*)$ is the unique minimizer. The consistency then follows from the argmax continuous mapping theorem under Assumption (A7).

Denote $K_1 = \varphi_0(\boldsymbol{X}) - \varphi(\boldsymbol{X};\boldsymbol{\gamma}) + \epsilon$, $K_2 = (A - \pi(\boldsymbol{X}))\Delta C(\boldsymbol{X};\boldsymbol{\beta})$. By Lemma 2,

$$\begin{aligned} D = &-\tau K_2 1\{K_1 \geq 0\} + (1-\tau)K_2 1\{K_1 < 0\} + (K_2 - K_1)1\{K_1 \geq 0\}1\{K_2 > K_1\} \\ &+ (K_1 - K_2)1\{K_1 < 0\}1\{K_2 < K_1\}. \end{aligned}$$

Since $\epsilon \perp A|\boldsymbol{X}$ and $\Pr(A|\boldsymbol{X}) = \pi(\boldsymbol{X})$, applying double expectation rule with $\boldsymbol{X}$, we have $\mathrm{E}[-\tau K_2 1\{K_1 \geq 0\}] = \mathrm{E}[(1-\tau)K_2 1\{K_1 < 0\}] = 0$. Thus,

$$G_1(\boldsymbol{\beta},\boldsymbol{\gamma}) = \mathrm{E}[(K_2 - K_1)1\{K_1 \geq 0\}1\{K_2 > K_1\}] + \mathrm{E}[(K_1 - K_2)1\{K_1 < 0\}1\{K_2 < K_1\}]. \quad (4.17)$$

It is easy to check $G_1(\boldsymbol{\beta},\boldsymbol{\gamma}) \geq 0$ and achieves minimal value 0 at point $(\boldsymbol{\beta}_0,\boldsymbol{\gamma})$ for all $\boldsymbol{\gamma} \in \Gamma$. In addition, by A6, we know $G_2(\boldsymbol{\gamma})$ has unique minimizer $\boldsymbol{\gamma}^*$. Combining the above two facts, a) is proved.

Combining A4, A8 and (4.17), we could prove $G_1(\boldsymbol{\beta},\boldsymbol{\gamma}) > 0$ for all $\boldsymbol{\beta} \neq \boldsymbol{\beta}_0$ and $\boldsymbol{\gamma} \in \Gamma$. So b) holds. $\square$

Proof of Theorem 4.2. (a) When $M(x) = H_\alpha(x)$, the proof follows similar steps as Theorem 1. The only difference is that $G_1(\boldsymbol{\beta},\boldsymbol{\gamma})$ takes a different expression now and we need to redo the proof of 1) $G_1(\boldsymbol{\beta},\boldsymbol{\gamma}) > 0$ $\forall \boldsymbol{\beta} \neq \boldsymbol{\beta}_0$, $\boldsymbol{\gamma} \in \Gamma$, and 2) $G_1(\boldsymbol{\beta}_0,\boldsymbol{\gamma}) = 0$ $\forall \boldsymbol{\gamma} \in \Gamma$. By definition, $G_1(\boldsymbol{\beta},\boldsymbol{\gamma}) \triangleq \mathrm{E}(D)$, where

$$D = H_\alpha \left[\varphi_0(\boldsymbol{X}) - \varphi(\boldsymbol{X};\boldsymbol{\gamma}) + \epsilon - (A - \pi(\boldsymbol{X}))\Delta C(\boldsymbol{X};\boldsymbol{\beta})\right] - H_\alpha \left[\varphi_0(\boldsymbol{X}) - \varphi(\boldsymbol{X};\boldsymbol{\gamma}) + \epsilon\right].$$

Then, 2) holds immediately. Denote $K_1 = \varphi_0(\boldsymbol{X}) - \varphi(\boldsymbol{X};\boldsymbol{\gamma}) + \epsilon$, $K_2 = (A - \pi(\boldsymbol{X}))\Delta C(\boldsymbol{X};\boldsymbol{\beta})$. We have the following four cases:

1. If $K_1 > \alpha$ then $H_\alpha(K_1 - K_2) \geq \alpha(K_1 - K_2) - 0.5\alpha^2$. Thus, $D \geq -\alpha K_2$;

2. If $K_1 < -\alpha$ then $H_\alpha(K_1 - K_2) \geq \alpha(K_2 - K_1) - 0.5\alpha^2$. Thus, $D \geq \alpha K_2$;

3. If $K_1 \in [-\alpha, \alpha]$ and $K_1 - K_2 \in [-\alpha, \alpha]$ then $D = 1/2(K_1 - K_2)^2 - 1/2K_1^2 = -K_1K_2 + 1/2K_2^2$;

4. If $K_1 \in [-\alpha, \alpha]$ and $K_1 - K_2 \notin [-\alpha, \alpha]$ then

$$\begin{aligned} &H_\alpha(K_1 - K_2) \geq 1/2(K_1 - K_2)^2 - \left\{1/2(\alpha + |K_2|)^2 - \left[\alpha(\alpha + |K_2|) - 1/2\alpha^2\right]\right\} \\ =&1/2(K_1 - K_2)^2 - 1/2K_1^2. \end{aligned}$$

   Thus, $D \geq 1/2(K_1 - K_2)^2 - 1/2K_1^2 - 1/2K_2^2 = -K_1K_2$.

Combining the above four equalities and inequalities,

$$\begin{aligned} G_1(\boldsymbol{\beta}, \boldsymbol{\gamma}) \geq &\mathrm{E}[-\alpha K_2 1\{K_1 > \alpha\}] + \mathrm{E}[\alpha K_2 1\{K_1 < -\alpha\}] + \mathrm{E}[-K_1K_2 1\{K_1 \in [-\alpha, \alpha]\}] \\ &+ \mathrm{E}\left[1/2K_2^2 1\left(\{K_1 \in [-\alpha, \alpha]\} \cup \{K_1 - K_2 \in [-\alpha, \alpha]\}\right)\right] \end{aligned}$$

Since $\epsilon \perp A|\boldsymbol{X}$ and $\Pr(A|\boldsymbol{X}) = \pi(\boldsymbol{X})$, applying double expectation rule with $\boldsymbol{X}$, we have $\mathrm{E}[-\alpha K_2 1\{K_1 > \alpha\}] = \mathrm{E}[\alpha K_2 1\{K_1 < -\alpha\}] = \mathrm{E}[-K_1K_2 1\{K_1 \in [-\alpha, \alpha]\}] = 0$. Thus,

$$G_1(\boldsymbol{\beta}; \boldsymbol{\gamma}) \geq \mathrm{E}\left[1/2K_2^2 1\left(\{K_1 \in [-\alpha, \alpha]\} \cup \{K_1 - K_2 \in [-\alpha, \alpha]\}\right)\right]. \tag{4.18}$$

Combining (4.18), A4 and A8, we can check that 1) holds. Thus, part (a) is proved.

(b) When $M(x) = J_\epsilon(x)$, similarly $D = J_\epsilon(K_1 - K_2) - J_\epsilon(K_1)$. Notice that we have the following three cases:

1. If $K_1 > \epsilon$ then $D \geq -K_2$;

2. If $K_1 < -\epsilon$ then $D \geq K_2$;

3. If $K_1 \in [-\epsilon, \epsilon]$ then $D \geq 0$;

The rest of the proof follows similar steps as part (a). $\square$

Proof of Theorem 4.5. From Theorem 4.1, $\boldsymbol{\beta}_\tau = \boldsymbol{\beta}_0$. Plugging this into Theorem 4.4 and applying double expectation rules, we have

$$J(\tau) = \mathrm{E}\left[f_\epsilon\left(\tilde{\boldsymbol{X}}^\intercal \boldsymbol{\gamma}(\tau) - \varphi_0(\boldsymbol{X})|\boldsymbol{X}\right)\begin{pmatrix} \pi(\boldsymbol{X})(1-\pi(\boldsymbol{X}))\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\intercal & \mathbf{0} \\ \mathbf{0} & \tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\intercal \end{pmatrix}\right]$$

and

$$\Sigma(\tau,\tau) = \mathrm{E}\left\{\left[\tau - 1\left\{\epsilon < \tilde{\boldsymbol{X}}^\intercal \boldsymbol{\gamma}(\tau) - \varphi_0(\boldsymbol{X})\right\}\right]^2 \begin{pmatrix} \pi(\boldsymbol{X})(1-\pi(\boldsymbol{X}))\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\intercal & \mathbf{0} \\ \mathbf{0} & \tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\intercal \end{pmatrix}\right\}.$$

Thus, $\sqrt{n}(\hat{\boldsymbol{\beta}}(\tau) - \boldsymbol{\beta}_0) \xrightarrow{d} N(\mathbf{0}, J_{11}^{-1}(\tau)\Sigma_{11}(\tau,\tau)J_{11}^{-1}(\tau))$, where $J_{11}^{-1}(\tau)$ and $\Sigma_{11}(\tau,\tau)$ are defined as in Theorem 5. Conditional on $\boldsymbol{X}$, $1\left\{\epsilon < \tilde{\boldsymbol{X}}^\intercal \boldsymbol{\gamma}(\tau) - \varphi_0(\boldsymbol{X})\right\}$ is a binomial random variable with $p = \Pr\left(\epsilon < \tilde{\boldsymbol{X}}^\intercal \boldsymbol{\gamma}(\tau) - \varphi_0(\boldsymbol{X})\right)$. Then, $\mathrm{E}\left\{\left[\tau - 1\{\epsilon < \tilde{\boldsymbol{X}}^\intercal \boldsymbol{\gamma}(\tau) - \varphi_0(\boldsymbol{X})\}\right]^2 |\boldsymbol{X}\right\} = (p - \tau)^2 + p(1-p) \leq \tau^2 + |1-2\tau|$. Thus, $\Sigma_{11}(\tau,\tau) \leq \left(\tau^2 + |1-2\tau|\right)\mathrm{E}\left[\pi(\boldsymbol{X})(1-\pi(\boldsymbol{X}))\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\intercal\right]$. $\square$

4.7.2 Additional Simulation Result

Table 4.6: Summary result of Model I with non-constant propensity scores. $\delta_{0.5}$ is multiplied by 10.

		Homogeneous Error								
		Normal			**Log-Normal**			**Cauchy**		
n	method	mse	PCD	$\delta_{0.5}$	mse	PCD	$\delta_{0.5}$	mse	PCD	$\delta_{0.5}$
100	LS	1.70 (0.061)	81.9	0.91	2.90 (0.114)	77.6	1.34		59.3	3.61
	P(0.5)	1.90 (0.069)	80.1	1.09	2.13 (0.073)	78.3	1.25	3.54 (0.128)	75.7	1.57
	P(0.25)	2.35 (0.080)	78.2	1.33	1.95 (0.076)	80.4	1.08	8.45 (0.431)	69.8	2.28
	Huber	1.51 (0.053)	82.1	0.89	1.77 (0.065)	80.6	1.02	3.67 (0.127)	75.4	1.60
200	LS	0.77 (0.026)	86.8	0.50	1.35 (0.045)	82.2	0.91		59.2	3.63
	P(0.5)	0.88 (0.028)	85.5	0.60	1.00 (0.029)	83.0	0.79	1.54 (0.050)	81.1	1.00
	P(0.25)	1.06 (0.035)	84.5	0.68	0.83 (0.027)	85.9	0.59	3.61 (0.143)	74.7	1.70
	Huber	0.68 (0.022)	87.3	0.46	0.81 (0.025)	85.2	0.62	1.58 (0.052)	80.7	1.03
400	LS	0.39 (0.012)	90.2	0.28	0.65 (0.020)	86.9	0.48		58.0	3.79
	P(0.5)	0.43 (0.013)	89.3	0.32	0.47 (0.014)	88.4	0.38	0.73 (0.022)	86.5	0.51
	P(0.25)	0.53 (0.016)	88.5	0.38	0.41 (0.013)	90.5	0.27	1.50 (0.049)	81.7	0.96
	Huber	0.34 (0.010)	90.6	0.25	0.39 (0.012)	89.6	0.30	0.72 (0.022)	86.3	0.53
800	LS	0.18 (0.006)	93.3	0.13	0.32 (0.010)	90.2	0.27		58.3	3.75
	P(0.5)	0.21 (0.007)	92.7	0.15	0.24 (0.007)	91.5	0.20	0.36 (0.011)	90.3	0.27
	P(0.25)	0.28 (0.009)	92.4	0.17	0.21 (0.007)	93.4	0.13	0.78 (0.026)	86.9	0.50
	Huber	0.16 (0.005)	93.7	0.11	0.19 (0.006)	92.6	0.15	0.37 (0.010)	89.9	0.28
		Heterogenous Error								
		Normal			**Log-Normal**			**Cauchy**		
n	method	mse	PCD	$\delta_{0.5}$	mse	PCD	$\delta_{0.5}$	mse	PCD	$\delta_{0.5}$
100	LS	2.84 (0.111)	78.2	1.33	9.96 (0.773)	72.0	2.06		55.2	4.18
	P(0.5)	2.01 (0.082)	80.6	1.09	2.18 (0.080)	79.2	1.21	4.18 (0.189)	74.1	1.81
	P(0.25)	2.91 (0.110)	76.7	1.52	3.22 (0.105)	74.2	1.76	10.62 (0.475)	65.3	2.87
	Huber	1.90 (0.074)	80.9	1.06	2.38 (0.090)	78.1	1.32	5.06 (0.230)	71.9	2.04
200	LS	1.46 (0.053)	83.1	0.83	4.47 (0.371)	76.8	1.51		56.3	4.04
	P(0.5)	0.92 (0.033)	86.4	0.55	0.98 (0.035)	85.3	0.64	1.69 (0.065)	81.5	0.98
	P(0.25)	1.35 (0.049)	83.3	0.81	1.47 (0.049)	81.6	0.97	4.73 (0.241)	71.9	2.05
	Huber	0.86 (0.030)	86.6	0.52	1.02 (0.036)	84.7	0.68	2.11 (0.079)	79.3	1.18
400	LS	0.74 (0.029)	87.4	0.47	2.65 (0.402)	81.4	1.04		56.2	4.06
	P(0.5)	0.45 (0.016)	90.2	0.29	0.44 (0.017)	89.5	0.34	0.79 (0.029)	87.2	0.49
	P(0.25)	0.66 (0.025)	88.3	0.41	0.70 (0.023)	86.9	0.50	2.12 (0.091)	79.5	1.19
	Huber	0.43 (0.016)	90.2	0.28	0.48 (0.018)	89.0	0.36	1.01 (0.036)	85.0	0.65
800	LS	0.36 (0.013)	90.8	0.25	1.09 (0.066)	85.0	0.69		56.3	4.02
	P(0.5)	0.21 (0.008)	93.2	0.14	0.24 (0.009)	92.3	0.19	0.39 (0.014)	90.5	0.27
	P(0.25)	0.33 (0.013)	91.7	0.21	0.36 (0.012)	90.8	0.25	1.01 (0.034)	84.9	0.65
	Huber	0.20 (0.008)	93.2	0.14	0.25 (0.009)	92.1	0.19	0.49 (0.016)	89.1	0.34

Table 4.7: Summary result of Model II with non-constant propensity scores. $\delta_{0.5}$ is multiplied by 10.

		Homogeneous Error								
		Normal			**Log-Normal**			**Cauchy**		
n	method	mse	PCD	$\delta_{0.5}$	mse	PCD	$\delta_{0.5}$	mse	PCD	$\delta_{0.5}$
100	LS	0.36 (0.011)	89.8	0.29	1.65 (0.085)	80.8	1.06		58.7	3.69
	P(0.5)	0.57 (0.017)	86.9	0.46	0.61 (0.026)	86.4	0.55	1.31 (0.045)	81.7	0.93
	P(0.25)	0.65 (0.020)	86.2	0.52	0.22 (0.008)	91.7	0.20	4.67 (0.312)	74.7	1.64
	Huber	0.38 (0.012)	89.5	0.30	0.45 (0.018)	88.3	0.40	1.70 (0.060)	79.5	1.14
200	LS	0.16 (0.004)	92.9	0.14	0.74 (0.030)	85.6	0.61		59.1	3.64
	P(0.5)	0.25 (0.007)	91.2	0.21	0.26 (0.008)	90.7	0.24	0.52 (0.017)	87.8	0.41
	P(0.25)	0.30 (0.008)	90.3	0.26	0.09 (0.003)	94.8	0.08	1.69 (0.074)	81.3	0.92
	Huber	0.17 (0.005)	92.8	0.14	0.19 (0.006)	92.2	0.17	0.70 (0.022)	86.2	0.53
400	LS	0.08 (0.002)	95.1	0.06	0.36 (0.013)	89.7	0.30		58.0	3.79
	P(0.5)	0.12 (0.003)	93.8	0.10	0.12 (0.003)	93.8	0.10	0.22 (0.006)	91.6	0.19
	P(0.25)	0.14 (0.004)	93.3	0.12	0.04 (0.001)	96.5	0.03	0.63 (0.021)	86.5	0.49
	Huber	0.08 (0.002)	95.0	0.07	0.09 (0.002)	94.8	0.07	0.30 (0.009)	90.3	0.26
800	LS	0.04 (0.001)	96.5	0.03	0.18 (0.006)	92.3	0.16		58.2	3.76
	P(0.5)	0.06 (0.002)	95.6	0.05	0.06 (0.002)	95.6	0.05	0.10 (0.003)	94.4	0.09
	P(0.25)	0.07 (0.002)	95.3	0.06	0.02 (0.001)	97.5	0.02	0.29 (0.009)	90.6	0.23
	Huber	0.04 (0.001)	96.4	0.03	0.04 (0.001)	96.3	0.04	0.14 (0.004)	93.2	0.12
		Heterogenous Error								
		Normal			**Log-Normal**			**Cauchy**		
n	method	mse	PCD	$\delta_{0.5}$	mse	PCD	$\delta_{0.5}$	mse	PCD	$\delta_{0.5}$
100	LS	1.45 (0.059)	82.9	0.85	8.53 (0.784)	72.4	2.01		54.9	4.22
	P(0.5)	0.94 (0.034)	85.6	0.61	1.29 (0.058)	83.3	0.86	2.27 (0.132)	78.9	1.24
	P(0.25)	1.46 (0.051)	81.5	0.96	1.78 (0.071)	78.2	1.30	7.88 (0.422)	68.1	2.46
	Huber	0.89 (0.034)	86.1	0.57	1.46 (0.067)	81.7	0.99	3.28 (0.157)	75.1	1.65
200	LS	0.84 (0.035)	86.6	0.53	3.85 (0.358)	77.6	1.43		55.9	4.09
	P(0.5)	0.44 (0.016)	90.0	0.29	0.60 (0.024)	89.0	0.39	0.87 (0.034)	86.3	0.56
	P(0.25)	0.69 (0.025)	87.0	0.49	0.75 (0.024)	85.5	0.59	3.08 (0.179)	75.3	1.58
	Huber	0.43 (0.016)	90.3	0.28	0.66 (0.025)	87.7	0.47	1.32 (0.050)	82.4	0.87
400	LS	0.44 (0.020)	90.3	0.28	2.34 (0.393)	82.4	0.95		55.9	4.09
	P(0.5)	0.23 (0.009)	92.9	0.16	0.28 (0.011)	92.5	0.19	0.39 (0.015)	90.8	0.26
	P(0.25)	0.33 (0.011)	91.0	0.23	0.36 (0.012)	90.1	0.27	1.25 (0.048)	82.8	0.82
	Huber	0.22 (0.008)	93.1	0.15	0.31 (0.012)	91.7	0.21	0.60 (0.022)	88.0	0.43
800	LS	0.23 (0.009)	93.0	0.15	0.90 (0.057)	86.2	0.60		56.3	4.03
	P(0.5)	0.11 (0.004)	95.0	0.07	0.14 (0.005)	94.8	0.09	0.18 (0.006)	93.6	0.12
	P(0.25)	0.17 (0.006)	93.7	0.12	0.18 (0.006)	93.0	0.14	0.59 (0.017)	87.3	0.44
	Huber	0.10 (0.004)	95.1	0.07	0.15 (0.006)	94.2	0.11	0.29 (0.010)	91.4	0.21

Table 4.8: Summary results with non-constant propensity scores when errors interacted with treatment.

			Least Square					**Pinball(0.5)**					**Pinball(0.25)**				
Error	d_0	n	mse	PCD	δ_μ	$\delta_{0.5}$	$\delta_{0.25}$	mse	PCD	δ_μ	$\delta_{0.5}$	$\delta_{0.25}$	mse	PCD	δ_μ	$\delta_{0.5}$	$\delta_{0.25}$
Normal	0	100	3.90	87.8	0.06	0.06	0.73	6.33	84.5	0.09	0.09	0.76	7.98	82.9	0.10	0.10	0.78
		200	1.78	91.7	0.03	0.03	0.70	2.83	89.5	0.04	0.04	0.72	3.68	88.2	0.05	0.05	0.72
		400	0.75	94.4	0.01	0.01	0.69	1.26	92.8	0.02	0.02	0.69	1.70	91.7	0.02	0.02	0.70
		800	0.38	96.0	0.01	0.01	0.68	0.63	94.8	0.01	0.01	0.68	0.84	94.1	0.01	0.01	0.69
	5	100	18.97	76.4	0.19	0.19	1.31	15.58	75.2	0.20	0.20	1.30	21.14	73.4	0.30	0.30	1.19
		200	8.40	82.4	0.12	0.12	1.24	7.39	80.4	0.14	0.14	1.22	11.36	80.4	0.22	0.22	1.08
		400	3.97	87.8	0.06	0.06	1.18	4.22	84.9	0.09	0.09	1.14	7.35	85.3	0.17	0.17	1.02
		800	1.98	91.7	0.03	0.03	1.13	2.93	87.8	0.06	0.06	1.08	5.43	88.0	0.15	0.15	1.00
	10	100	49.37	68.6	0.30	0.30	1.76	23.68	70.6	0.26	0.26	1.71	48.69	67.9	0.43	0.43	1.32
		200	22.74	74.7	0.21	0.21	1.72	11.37	76.3	0.18	0.18	1.61	31.67	74.5	0.37	0.37	1.14
		400	11.09	80.8	0.14	0.14	1.66	6.86	80.5	0.14	0.14	1.51	24.55	78.2	0.31	0.31	1.05
		800	5.36	86.0	0.08	0.08	1.60	4.77	83.3	0.10	0.10	1.42	22.29	80.3	0.28	0.28	1.01
Gamma	0	100	4.33	87.3	0.06	0.06	0.06	5.97	84.8	0.09	0.09	0.09	5.61	85.9	0.08	0.08	0.08
		200	1.76	91.8	0.03	0.03	0.03	2.51	89.9	0.04	0.04	0.04	2.52	90.3	0.04	0.04	0.04
		400	0.79	94.2	0.01	0.01	0.01	1.18	93.0	0.02	0.02	0.02	1.18	93.1	0.02	0.02	0.02
		800	0.39	95.9	0.01	0.01	0.01	0.61	94.9	0.01	0.01	0.01	0.58	94.9	0.01	0.01	0.01
	5	100	18.45	78.1	0.17	0.20	0.34	10.89	77.8	0.20	0.17	0.24	11.60	79.1	0.27	0.17	0.15
		200	8.01	83.8	0.10	0.14	0.28	5.50	83.0	0.13	0.10	0.15	6.18	85.5	0.21	0.11	0.06
		400	4.00	88.0	0.06	0.09	0.23	3.45	86.9	0.09	0.05	0.09	4.53	88.2	0.18	0.08	0.04
		800	1.93	91.9	0.03	0.05	0.19	2.55	88.8	0.08	0.03	0.06	3.48	89.5	0.16	0.06	0.03
	10	100	49.62	71.0	0.27	0.35	0.92	20.82	74.5	0.28	0.24	0.65	33.03	75.0	0.46	0.20	0.32
		200	23.14	76.0	0.19	0.32	0.94	11.48	81.0	0.20	0.15	0.55	24.14	79.3	0.42	0.13	0.22
		400	11.46	81.0	0.13	0.26	0.90	8.23	85.3	0.16	0.09	0.46	20.72	82.0	0.39	0.10	0.17
		800	5.50	86.4	0.08	0.20	0.83	6.68	88.1	0.13	0.05	0.40	19.38	82.9	0.36	0.07	0.16

REFERENCES

Ahmad, I., Leelahanon, S., and Li, Q. (2005). Efficient estimation of a semiparametric partially linear varying coefficient model. *Ann. Statist.*, 33(1):258–283.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory (Tsahkadsor, 1971)*, pages 267–281. Akadémiai Kiadó, Budapest.

Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: a large sample study. *Ann. Statist.*, 10(4):1100–1120.

Angrist, J., Chernozhukov, V., and Fernández-Val, I. (2006). Quantile regression under misspecification, with an application to the us wage structure. *Econometrica*, 74(2):539–563.

Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics*, 37(4):pp. 373–384.

Cai, Z., Fan, J., and Li, R. (2000). Efficient estimation and inferences for varying-coefficient models. *Journal of the American Statistical Association*, 95(451):pp. 888–902.

Cai, Z. and Sun, Y. (2003). Local linear estimation for time-dependent coefficients in Cox's regression models. *Scand. J. Statist.*, 30(1):93–111.

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.

Cox, D. R. (1972). Regression models and life-tables. *J. Roy. Statist. Soc. Ser. B*, 34:187–220. With discussion by F. Downton, Richard Peto, D. J. Bartholomew, D. V. Lindley, P. W. Glassborow, D. E. Barton, Susannah Howard, B. Benjamin, John J. Gart, L. D. Meshalkin, A. R. Kagan, M. Zelen, R. E. Barlow, Jack Kalbfleisch, R. L. Prentice and Norman Breslow, and a reply by D. R. Cox.

Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(425-455).

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *Ann. Statist.*, 32(2):407–499. With discussion, and a rejoinder by the authors.

Fan, J. and Huang, T. (2005). Profile likelihood inferences on semiparametric varying-coefficient partially linear models. *Bernoulli*, 11(6):1031–1057.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.*, 96(456):1348–1360.

Fan, J. and Li, R. (2002). Variable selection for Cox's proportional hazards model and frailty model. *Ann. Statist.*, 30(1):74–99.

Fleming, T. R. and Harrington, D. P. (1991). *Counting processes and survival analysis.* Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons Inc., New York.

Frank, A. and Asuncion, A. (2010). UCI machine learning repository.

Friedman, J. (2008). Fast sparse regression and classification.

Goeman, J. J. (2010). L1 penalized estimation in the cox proportional hazards model. *Biometrical Journal*, 52(1):70–84.

Hahn, J. (1997). Bayesian bootstrap of the quantile regression estimator: a large sample study. *International Economic Review*, pages 795–808.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer Series in Statistics. Springer, New York, second edition.

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

Hu, T. and Xia, Y. (2012). Adaptive semi-varying coefficient model selection. *Statist. Sinica*, 22(2):575–599.

Huang, J. Z., Wu, C. O., and Zhou, L. (2002). Varying-coefficient models and basis function approximations for the analysis of repeated measurements. *Biometrika*, 89(1):111–128.

Huang, T.-M. (2010). Testing conditional independence using maximal nonlinear conditional correlation. *The Annals of Statistics*, 38(4):2047–2091.

Kim, M.-O. (2007). Quantile regression with varying coefficients. *Ann. Statist.*, 35(1):92–108.

Koenker, R. (2005). *Quantile regression*, volume 38 of *Econometric Society Monographs*. Cambridge University Press, Cambridge.

Lange, K. (2004). *Optimization*. Springer Texts in Statistics. Springer-Verlag, New York.

Lawrance, A. (1976). On conditional and partial correlation. *The American Statistician*, 30(3):146–149.

Ledoux, M. and Talagrand, M. (1991). *Probability in Banach spaces*, volume 23 of *Ergebnisse der Mathematik und ihrer Grenzgebiete (3) [Results in Mathematics and Related Areas (3)]*. Springer-Verlag, Berlin. Isoperimetry and processes.

Lee, Y.-Y. (2013). Interpretation and semiparametric efficiency in quantile regression under misspecification. *Working paper, https://sites.google.com/site/yyleelilian/*.

Leon, S., Tsiatis, A. A., and Davidian, M. (2003). Semiparametric estimation of treatment effect in a pretest-posttest study. *Biometrics*, 59(4):1046–1055.

Li, G., Peng, H., and Zhu, L. (2011). Nonconcave penalized m-estimation with a diverging number of parameters. *Statistica Sinica*, 21(1):391.

Li, R. and Liang, H. (2008). Variable selection in semiparametric regression modeling. *Ann. Statist.*, 36(1):261–286.

Liu, M., Lu, W., Krogh, V., Hallmans, G., Clendenen, T. V., and Zeleniuch-Jacquotte, A. (2013). Estimation and selection of complex covariate effects in pooled nested case–control studies with heterogeneity. *Biostatistics*, 14(4):682–694.

Lu, W., Zhang, H. H., and Zeng, D. (2013). Variable selection for optimal treatment decision. *Statistical methods in medical research*, 22(5):493–504.

Mallows, C. L. (1973). Some comments on cp. *Technometrics*, 15(4):661–675.

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979). *Multivariate Analysis*. Academic Press [Harcourt Brace Jovanovich Publishers], London. Probability and Mathematical Statistics: A Series of Monographs and Textbooks.

Meier, L., van de Geer, S., and Bühlmann, P. (2008). The group Lasso for logistic regression. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 70(1):53–71.

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.

Neyman, J. (1923). Sur les applications de la thar des probabilities aux experiences agaricales: Essay des principle. excerpts reprinted (1990) in english. *Statistical Science*, 5:463–472.

Ortega, J. M. and Rheinboldt, W. C. (2000). *Iterative Solution of Nonlinear Equations in Several Variables*, volume 30 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA. Reprint of the 1970 original.

Park, M. Y. and Hastie, T. (2006). Regularization path algorithms for detecting gene interactions. Technical report, Stanford University.

Qian, M. and Murphy, S. A. (2011). Performance guarantees for individualized treatment rules. *Ann. Statist.*, 39(2):1180–1210.

Robins, J. M. (2000). Marginal structural models versus structural nested models as tools for causal inference. In *Statistical models in epidemiology, the environment, and clinical trials (Minneapolis, MN, 1997)*, volume 116 of *IMA Vol. Math. Appl.*, pages 95–133. Springer, New York.

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.

Rosset, S. and Zhu, J. (2007). Piecewise linear regularized solution paths. *Annals of Statistics*, 35:1012–1030.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688.

Rubin, D. B. (1986). Comment: Which ifs have causal answers. *Journal of the American Statistical Association*, 81(396):961–962.

Schulte, P. J., Tsiatis, A. A., Laber, E. B., and Davidian, M. (2012). Q-and a-learning methods for estimating optimal dynamic treatment regimes. *arXiv preprint arXiv:1202.4177*.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.*, 6(2):461–464.

Song, K. (2009). Testing conditional independence via Rosenblatt transforms. *Ann. Statist.*, 37(6B):4011–4045.

Su, L. and White, H. (2007). A consistent characteristic function-based test for conditional independence. *Journal of Econometrics*, 141(2):807–834.

Sun, J. and Wei, L. J. (2000). Regression analysis of panel count data with covariate-dependent observation and censoring times. *Journal of The Royal Statistical Society Series B*, 62(2):293–302.

Tian, L., Zucker, D., and Wei, L. J. (2005). On the Cox model with time-varying regression coefficients. *J. Amer. Statist. Assoc.*, 100(469):172–183.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288.

Tibshirani, R. (1997). The lasso method for variable selection in the cox model. *Statistics in Medicine*, 16(4):385–395.

Tong, X., He, X., Sun, L., and Sun, J. (2009a). Variable selection for panel count data via non-concave penalized estimating function. *Scandinavian Journal of Statistics*, 36(4):620–635.

Tong, X., Zhu, L., and Sun, J. (2009b). Variable selection for recurrent event data via nonconcave penalized estimating function. *Lifetime Data Analysis*, 15(2):197–215.

Tsiatis, A. A., Davidian, M., Zhang, M., and Lu, X. (2008). Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: A principled yet flexible approach. *S-tatistics in medicine*, 27(23):4658–4677.

Wang, H. and Xia, Y. (2009). Shrinkage estimation of the varying coefficient model. *J. Amer. Statist. Assoc.*, 104(486):747–757.

Wang, H. J., Zhu, Z., and Zhou, J. (2009). Quantile regression in partially linear varying coefficient models. *Ann. Statist.*, 37(6B):3841–3866.

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

Wu, Y. (2011). An ordinary differential equation-based solution path algorithm. *Journal of Nonparametric Statistics*, 23:185–199.

Wu, Y. (2012). Elastic net for Coxs proportional hazards model with a solution path algorithm. *Statistical Sinica*, 22:271–294.

Wu, Y. and Liu, Y. (2009). Variable selection in quantile regression. *Statistica Sinica*, 19(2):801.

Yan, J. and Huang, J. (2012). Model selection for Cox models with time-varying coefficients. *Biometrics*, 68(2):419–428.

Yandell, B. S. (1983). Nonparametric inference for rates with censored survival data. *Ann. Statist.*, 11(4):1119–1135.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 68(1):49–67.

Yuan, M. and Lin, Y. (2007). On the non-negative garrotte estimator. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):143–161.

Zhang, B., Tsiatis, A. A., Laber, E. B., and Davidian, M. (2013). Robust estimation of optimal dynamic treatment regimes for sequential treatment decisions. *Biometrika*, 100(3):681–694.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.*, 38(2):894–942.

Zhang, H. H., Cheng, G., and Liu, Y. (2011). Linear or nonlinear? automatic structure discovery for partially linear models. *Journal of the American Statistical Association*, 106(495):1099–1112.

Zhang, H. H. and Lu, W. (2007). Adaptive Lasso for Cox's proportional hazards model. *Biometrika*, 94(3):691–703.

Zhang, K., Peters, J., Janzing, D., and Schölkopf, B. (2012). Kernel-based conditional independence test and application in causal discovery. *arXiv preprint arXiv:1202.3775*.

Zhang, M., Tsiatis, A. A., and Davidian, M. (2008). Improving efficiency of inferences in randomized clinical trials using auxiliary covariates. *Biometrics*, 64(3):707–715.

Zhang, W., Lee, S.-Y., and Song, X. (2002). Local polynomial fitting in semivarying coefficient model. *J. Multivariate Anal.*, 82(1):166–188.

Zhao, Y., Zeng, D., Rush, A. J., and Kosorok, M. R. (2012). Estimating individualized treatment rules using outcome weighted learning. *J. Amer. Statist. Assoc.*, 107(499):1106–1118.

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

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

Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Annals of statistics*, 36(4):1509.

Zucker, D. M. and Karr, A. F. (1990). Nonparametric survival analysis with time-dependent covariate effects: a penalized partial likelihood approach. *Ann. Statist.*, 18(1):329–353.