Quantile Regression Model Selection

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Benjamin Stanley Sherwood

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Lan Wang, Adviser

May 2014

Acknowledgements

I would like to thank my advisor, Lan Wang. I am very grateful for the time and energy she has spent to make this dissertation possible. She has been patient, while making sure I did not get off track. I have learned a great deal from working with Lan and her guidance has played a large role in grad school being a positive experience for me.

I would also like to thank the committee members Sandy Weisberg, Glen Meeden and Niels Waller. I appreciate them taking time to look at my work and respond with thoughtful comments.

I have very much enjoyed my time at the School of Statistics and a large part of that has been the great people I have met. I thank all the faculty and my fellow students in the School of Statistics.

I would like to thank my family. No grad school experience is complete without some crises. My parents and brother have all been supportive when support was needed. Most of all thanks to my wife, Sarah, for her support the past five years.

Abstract

Quantile regression models the conditional quantile of a response variable. Compared to least squares, which focuses on the conditional mean, it provides a more complete picture of the conditional distribution. Median regression, a special case of quantile regression, offers a robust alternative to least squares methods. Common regression assumptions are that there is a linear relationship between the covariates, there is no missing data and the sample size is larger than the number of covariates. In this dissertation we examine how to use quantile regression models when these assumptions do not hold. In all settings we examine the issue of variable selection and present methods that have the property of model selection consistency, that is, if the true model is one the candidate models, then these methods select the true model with probability approaching one as the sample size increases.

We consider partial linear models to relax the assumption that there is a linear relationship between the covariates. Partial linear models assume some covariates have a linear relationship with the response while other covariates have an unknown non-linear relationship. These models provide the flexibility of non-parametric methods while having ease of interpretation for the targeted parametric components. Additive partial linear models assume an additive form between the non-linear covariates, which allows for a flexible model that avoids the "curse of dimensionality". We examine additive partial linear quantile regression models using basis splines to model the non-linear relationships.

In practice missing data is a common problem and estimates can be biased if observations with missing data are dropped from the analysis. Imputation is a popular approach to handle missing data, but imputation methods typically require

distributional assumptions. An advantage of quantile regression is it does not require any distributional assumptions of the response or the covariates. To remain in a distribution free setting a different approach is needed. We use a weighted objective function that provides more weight to observations that are representative of subjects that are likely to have missing data. This approach is analyzed for both the linear and additive partial linear setting, while considering model selection for the linear covariates.

In mean regression analysis, detecting outliers and checking for non-constant variance are standard model-checking steps. With high-dimensional data, checking these conditions becomes increasingly cumbersome. Quantile regression offers an alternative that is robust to outliers in the Y direction and directly models heteroscedastic behavior. Penalized quantile regression is considered to accommodate models where the number of covariates is larger than the sample size. The additive partial linear model is extended to the high-dimensional case. We consider the setting where the number of linear covariates increases with the sample size, but the number of non-linear covariates remains fixed. To create a sparse model we compare the LASSO and SCAD penalties for the linear components.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Statisticians are often interested in modeling the relationship between a response variable and a set of covariates. Traditionally, this is achieved by least-squares type regression, which focuses on modeling the average of the response variable. However, in some important applications, non-central behavior of the response variable is of direct interest. For example, doctors may wish to identify the risk factors associated with low birth weights of infants. It is then natural to directly model the lower quantiles of the birth weight conditional on the covariates. It is noteworthy that the effect of a covariate may be different on the lower and higher quantiles of the response variable. For example, a certain treatment drug may be beneficial for relatively healthy patients, but could increase the risk of death for weaker patients. By considering different quantiles, we are able to obtain a more complete picture for the relationship of the covariates on the response variable.

Consider the research question of analyzing the relationship between X and Y after observing a random sample of $\{Y_i, x_{i1}, ..., x_{ip}\}_{i=1}^n$. Least squares regression can be used to model the conditional mean and, provided the errors are homoscedastic, the conditional variance. However, the mean and standard deviation can provide an incomplete description of the conditional distribution of $Y \mid X$. Often distributional assumptions such as $Y \mid X \sim N(g(X), \sigma^2)$ are made to simplify the modeling of

the conditional distributions, but estimates of conditional quantiles are sensitive to distributional assumptions. Another complication is the assumption that the variance is homoscedastic, which in many cases does not hold. Quantile regression avoids distributional assumptions by directly modeling the quantile of interest. If the researcher is only interested in central behavior quantile regression is a robust alternative to least squares based methods. Most importantly if the research question of interest is about a conditional quantile than we advocate estimating the conditional quantile directly by using quantile regression.

1.1 Structure

In the following chapters we provide a review of quantile regression, including current research in the area. In Chapter 2 the linear quantile regression model and corresponding objective function is presented. To understand how the objective function of quantile regression was derived we examine loss functions for unconditional quantiles. The quantile regression objective function is non-differentiable and we present the computational and theoretical challenges this provides. The linear quantile regression model is presented and compared to the typical linear mean regression model. In Chapter 3 we present the additive partial linear model using basis splines to model the non-linear variables. We present theorems stating that the non-linear estimation is consistent and that the linear components are asymptotically normal. Finite sample size performance of estimators are analyzed using Monte Carlo simulations.

In Chapter 4 we consider the problem of missing covariates in a quantile regression model. The three types of missingness: missing at random (MAR), missing completely at random (MCAR) and not missing at random (NMAR) are discussed along with a rationale for using the missing at random assumption. We provide mathematical intuition on why missing covariates can result in biased estimates. We propose an

inverse probability weighting (IPW) approach to provide unbiased estimates for both linear and additive partial linear models. Asymptotic results are stated along with Monte Carlo simulations and an analysis of health care cost data with missing data.

Chapter 5 examines additive partial linear quantile regression with a large number of linear covariates. When considering the asymptotics of these models we assume that the covariates grow with the sample size and allow for the number of covariates to be larger than the sample size. Sparsity is assumed to estimate these high dimensional models. Quantile regression allows for a nuanced definition of sparsity, it assumes that for a fixed τ, the quantile of interest, the model is sparse, but the active variables can change depending on τ. Our new contributions to this field are analyzing a high-dimensional additive partial linear model, where the number of linear covariates grows while the number of non-linear covariates remains fixed. Asymptotic results of the oracle model, the estimator we would use if we knew which linear covariates belonged in the model, are presented. For model selection of the linear terms we propose a penalized objective function. We are specifically interested in the SCAD penalty (Fan and Li, 2001) and demonstrate it has the oracle property in this setting.

We conclude with Chapter 6 which will discuss future research directions. In Chapter 4 we propose a weighting method, but it requires a correctly specified model for missingness. For mean regression double robust methods have been proposed that are robust to misspecification of the weighted model. For additive partial linear models we only considered model selection for the linear components. Future work could be done on model selection for the non-linear components using group penalties for the non-linear basis coefficients. Also, we could consider the setting where the number of non-linear variables increases with the sample size. Finally, our results have been limited to analyzing observations that are independent. It would be interesting to extend our results to handle longitudinal data.

1.2 Notation

Through out this document we use the following notation to notate different types of convergence:

1. $\xrightarrow{p}$ denotes convergence in probability,
2. $\xrightarrow{d}$ denotes convergence in distribution.

For any matrix A the spectral norm is used, that is,

$$||A|| = \sqrt{\lambda_{\max}(A'A)}.$$

Chapter 2

Linear Quantile Regression Model

2.1 Unconditional Estimation

Consider the continuous, random variable X with CDF $F(x)$ and $\mu = E[X]$. Define the τth quantile, $Q_\tau(X)$, as

$$Q_\tau(X) = \inf\{x : F(x) \geq \tau\}.$$

The median is the special case of $\tau = .5$ and we also use the notation of $\tilde{X} = Q_{.5}(X)$. If X has a finite second moment then

$$\mu = \underset{a}{\operatorname{argmin}}\, E(X - a)^2.$$

The population median minimizes the absolute error loss, that is

$$\tilde{X} = \underset{a}{\operatorname{argmin}}\, E|X - a|.$$

If we consider the *i.i.d.* sample of $X_1, \ldots, X_n$. Then the sample mean, $\bar{X}$, minimizes the squared error loss of the sample,

$$\bar{X} = \underset{a}{\operatorname{argmin}} \sum_{i=1}^{n} (X_i - a)^2.$$

The sample median, $\dot{X}$, may not be a unique value, but minimizing the absolute error loss provides a potential range of sample medians

$$\dot{X} = \underset{a}{\operatorname{argmin}} \sum_{i=1}^{n} |X_i - a|.$$

For a loss function approach to estimating the τth quantile we want a function $\rho_\tau(x - a)$ such that

$$Q_\tau(X) = \underset{a}{\operatorname{argmin}} \sum_{i=1}^{n} \rho_\tau(X_i - a).$$

The function that satisfies this is

$$\rho_\tau(u) = u(\tau - I(u < 0)).$$

The function $\rho_\tau(u)$ is called the check function because of it check shape as seen in Figure 2.1. Notice $\rho_{.5}(u) = .5|u|$ and for other values of τ the check function is a tilted absolute value function. An alternative definition of the check function is,

$$\rho_\tau(u) = \frac{1}{2}|u| + (\tau - 1/2)u.$$

Similar to how minimizing the squared or absolute error for single variable sample produces the sample mean or median, minimizing the check function error loss provides the corresponding sample quantile.

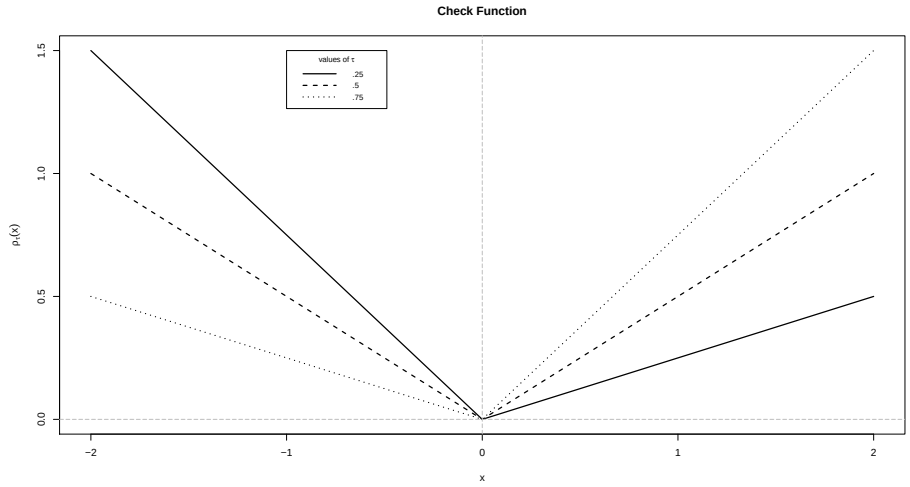

Figure 2.1: Check Function

2.2 Conditional Estimation

Now consider two random variables, $Y \in \mathcal{R}$ and $X \in \mathcal{R}^{p+1}$, including a constant. The conditional CDF of $Y \mid X$ is $F(y|X) = P(Y \leq y|X)$ then the conditional quantile is

$$Q_\tau\left(Y|X\right) = \inf\{y : F(y \mid X) \geq \tau\}.$$

Let $g(X)$ be any function of X and consider minimizing Y with respect to X.

$$\begin{aligned} E[Y \mid X] &= \underset{g(X)}{\operatorname{argmin}}\, E(Y - g(X))^2, \\ \text{median}[Y \mid X] &= \underset{g(X)}{\operatorname{argmin}}\, E|Y - g(X)|, \\ Q_\tau(Y \mid X) &= \underset{g(X)}{\operatorname{argmin}}\, E\rho_\tau(Y - g(X)). \end{aligned}$$

For an observed independent sample of $(Y_1, x_1), ..., (Y_n, x_n)$ where $x_i = (x_{i1}, ..., x_{ip})$ then the conditional mean could be estimated by

$$\hat{g}(X) = \underset{g(X)}{\operatorname{argmin}} \sum_{i=1}^{n} (Y_i - g(X_i))^2.$$

A common assumption used to derive an estimate $\hat{g}$ is to assume the linear model

$$\begin{aligned} Y_i &= \beta_{00} + \beta_{01} x_{i1} + ... + \beta_{0p} x_{ip} + \epsilon_i \\ &= x_i' \beta_0 + \epsilon_i, \end{aligned}$$

with $E[\epsilon_i] = 0$ and $\text{Var}(\epsilon_i) < \infty$. With the linear assumption estimating the conditional mean becomes a tractable problem with estimates of β_0 obtained by

$$\hat{\beta}(\mu) = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} (Y_i - x_i' \beta)^2. \tag{2.1}$$

For conditional quantiles we can consider a similar linear form for a fixed value of τ of

$$\begin{aligned} Y_i &= \beta_{00}(\tau) + \beta_{01}(\tau) x_{i1} + ... + \beta_{0p}(\tau) x_{ip} + \epsilon_i(\tau) \\ &= x_i' \beta_0(\tau) + \epsilon_i(\tau), \end{aligned} \tag{2.2}$$

with $P(\epsilon_i < 0 \mid x_i) = \tau$. No moment conditions on the error terms are required for quantile regression, providing insight that quantile regression outperforms mean regression if the error distribution is heavy-tailed. We have indexed $\beta_0(\tau)$ and $\epsilon_i(\tau)$ by τ because this model allows the relationship between the response and the covariates to change depending on the quantile of interest. This notation is too cumbersome and usually we drop the τ index, but it is important to understand that all quantile regression models discussed in this paper are for a fixed value of τ. Using this model

we can estimate the conditional quantile with

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} \rho_\tau(Y_i - x_i'\beta). \tag{2.3}$$

Minimizing (2.3) results in a quantile regression model. The progression of ideas that led to (2.3) motivated the original quantile regression model presented in Koenker and Bassett (1978).

2.3 Comparison to the Classic Linear Model

The classic linear model is

$$Y_i = x_i'\beta_0 + \epsilon_i,$$

with x_i and ϵ_i $i.i.d$, independent of each other and $\epsilon_i \sim N(0, \sigma^2)$. Then $Y_i \mid x_i \sim N(x_i'\beta_0, \sigma^2)$ and the conditional distribution of $Y_i \mid x_i$ can be approximated by estimating β_0 and σ^2 using OLS. In practice the assumption of homoscedastic, normally distributed errors are used to derive standard errors and p-values for $\hat{\beta}$. Under weaker conditions $\hat{\beta}$ is still a consistent and asymptotically normal estimator. Deviations from either of these assumptions implies the conditional distribution of $Y_i \mid X_i$ is not $N(x_i'\beta_0, \sigma^2)$ and therefore estimates of the conditional quantile are biased.

To compare quantile regression and least squares we consider a data set of monthly household income and food expenditure for 235 working class Belgian families collected by Engel (1857). Koenker and Bassett (1982) used the same data demonstrating how quantile regression can be used to test for heteroscedasticity in the data. They found that increases in family income increased both the average and variability of money spent on food. We use the same data, but remove a family with a

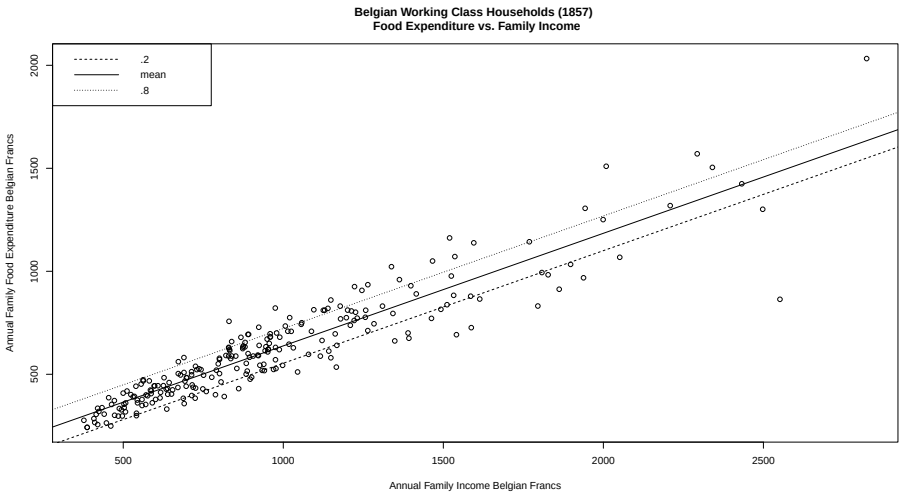

Figure 2.2: Least Squares Estimates of .2 and .8 Quantiles

household income of 4,957.8, while the next largest income was 2,822.5. While quantile regression is robust to outliers in the Y direction it can be influenced by outliers in the X direction. Figure 2.2 and Figure 2.3 are plots of the data with estimates for the conditional median and .2 and .8 quantiles. In Figure 2.2 the estimates are derived by using least squares method and assumptions of homoscedastic and normally distributed error terms. Modeling the .2 and .8 quantiles separately using the quantile regression objective function (2.3) provides the fits shown in Figure 2.3. The estimates shown in Figure 2.2 are problematic because too many lower income families are falling in between the .2 and .8 estimates and too many of the higher earning families are outside these estimates. The least squares based estimates of the conditional quantile are misspecified because the heteroscedastic relationship between food expenditure and income is not being modeled. The quantile regression estimates are able to model the heteroscedastic nature of the data. The slopes for the three different estimates, .2, .5 and .8 in Figure 2.3 have noticeably different slopes. Quantile regression provides a more flexible framework that allows the relationship between the response and the predictor to change depending on the quantile being modeled.

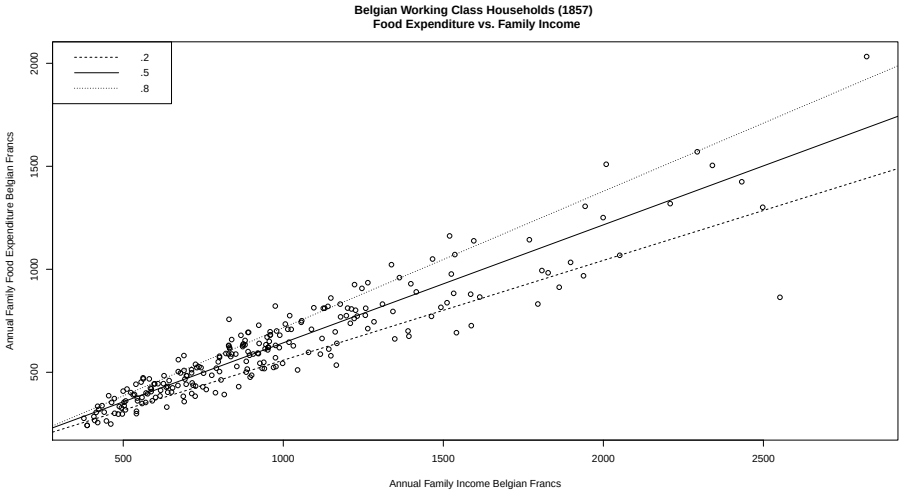

Figure 2.3: Quantile Regression Estimates of .2 and .8 Quantiles

To formally demonstrate this we consider the location-scale model

$$Y_i = x_i'\eta_0 + (x_i'\zeta_0)u_i, \tag{2.4}$$

where x_i is a vector of non-negative random variables and u_i are $i.i.d.$ mean zero random variables with CDF $F(\cdot)$ and inverse CDF $F^{-1}(\cdot)$. We require the elements of x_i to be non-negative to ensure the conditional quantiles have a linear relationship with the response. This is called the location-scale model because both the location and scale of the response varies with the covariates. Notice $E[Y_i \mid x_i] = x_i'\eta_0$ and $Q_\tau(Y_i \mid x_i) = x_i'\eta_0 + x_i'\zeta_0 F^{-1}(\tau)$.(Koenker, 2005) Focusing on the mean will only capture how the center of the response changes while ignoring the changes that occur to the scale of the model. The quantile regression coefficients for the τth quantile from (2.4) are $\beta_0(\tau) = \eta_0 + \zeta_0 F^{-1}(\tau)$. Therefore $\beta_0(\tau)$ depends on the scale, center and τ, while the conditional mean coefficients are only influenced by how covariates impact the expected value of the response. Even if the error terms are not heteroscedastic quantile regression is useful as a robust alternative to least squares methods.

Returning to Engel's data on spending patterns of working class families, Table 2.1 provides estimates, standard errors, t-statistics and p-values for the .2,.5 and .8 coefficients. Standard errors were calculated by bootstrapping on the families. Other methods exist for estimating standard errors, but are based on asymptotic distributions and require deciding if the error terms are independent or not.(Koenker, 2012) Using Table 2.1 we estimate that after accounting for income, that 80% of working class families spent 66% or less of their income on food, 50% spent 57% or less and 20% spent 48% or less of their income on food. Figure 2.4 plots the .05,.10,...,.90,.95 coefficients on the y-axis and τ on the x-axis. The black points represent a coefficient point estimate, the gray area represents 95% pointwise confidence lines, the middle line is the OLS fit and the dashed lines are the 95% confidence interval for the OLS estimate. If the covariate of interest only changes the location of the conditional distribution, but not the scale then the slope estimates should all be similar to the OLS estimate. In Figure 2.4 the estimates of $\beta(\tau)$ increase with τ which corresponds with what is seen in Figure 2.3, that both average and scale of food expenditure changes with income.

Tau	Income	St. Error	T-Value	P-Value T_{234}
0.20	0.48	0.02	20.34	0.00
0.50	0.57	0.03	20.16	0.00
0.80	0.66	0.03	25.12	0.00

Table 2.1: Engel Data Quantile Regression Coefficients

The location-scale model is helpful to understand the benefits of quantile regression, but it is actually a more specific model than quantile regression requires. Linear quantile regression only requires that (2.2) holds, where the key assumption is that the linear relationship holds for the quantile of interest. In Chapter 3 we consider an additive partial linear model which allows us to relax the assumption of a linear relationship, by assuming it holds for only a subset of the covariates. In Chapter 5 we

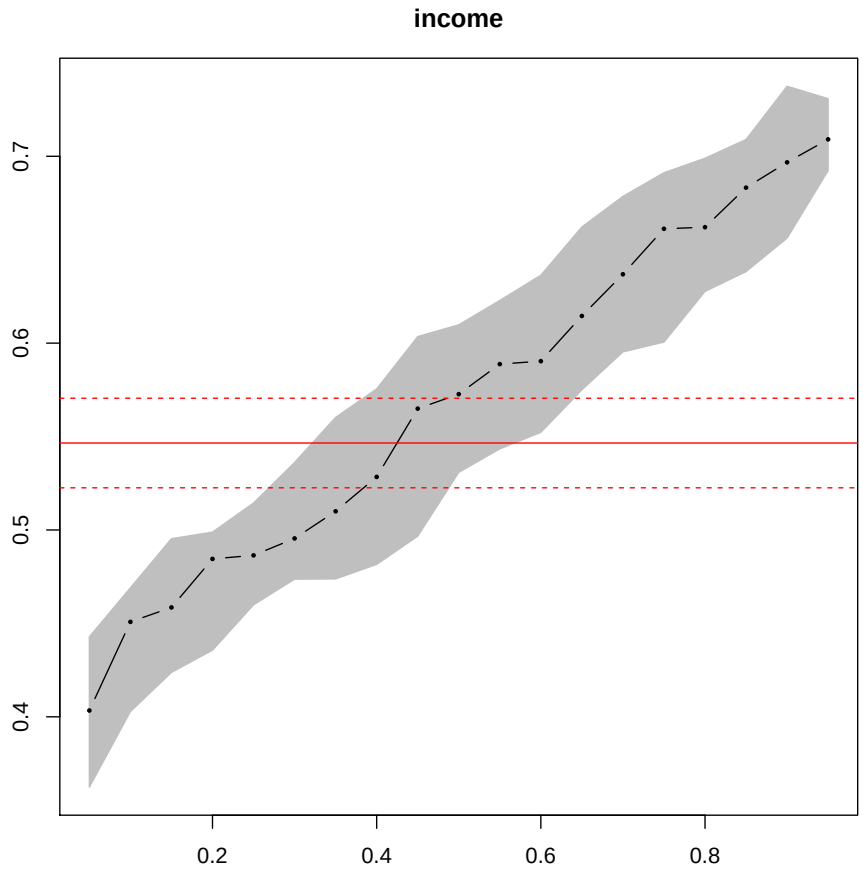

Figure 2.4: Engel Family Income Coefficients

examine quantile regression in high-dimensional setting and assume a sparse model to handle the case of $p >> n$.

A useful property of quantile regression, that will be used in some of our data analysis, is its equivariance to the monotone transformation of the response variable.(Koenker, 2005) More specifically, for any nondecreasing function $h(x)$

$$Q_{h(Y)}(\tau \mid X) = h(Q_Y(\tau \mid X)).$$

This can be derived from the fact $P(Y \leq y) = P(h(Y) \leq h(y))$. Mean regression does not share this property unless the transformation is linear. For this reason when interpreting transformed responses where the distribution of the error terms are symmetric, then interpretation on the original scale is actually for the median. The conditional median has the equivariance property, while the conditional mean is only equivariant under a linear transformation. On the other hand the linearity of the expectation operator is a nice property that is not shared by quantiles. This presents difficulties when considering model average estimates such as those derived from multiple imputation methods.(Wei et al., 2012)

2.4 Non-differentiability

2.4.1 Computational Difficulties

The quantile regression objective function is not differentiable which historically was a barrier to solving (2.3). It also provides challenges in understanding the asymptotic behavior of $\hat{\beta}$. For least squares choosing $\hat{\beta}(\mu)$ from (2.1) is equivalent to solving

$$\sum_{i=1}^{n} x_i(Y_i - x_i'\hat{\beta}(\mu)) = 0,$$

which implies $\hat{\beta}(\mu) = (X'X)^{-1}X'Y$. The objective function, $\rho_\tau(Y_i - x_i'\beta)$, is not differentiable and therefore another approach must be taken to solve the quantile regression objective function. The function $\sum_{i=1}^n \rho_\tau(Y_i - x_i'\beta)$ is differentiable except at points for which $Y_i - x_i'\beta = 0$. These points do have directional derivatives. Let $u \in \mathbb{R}^{p+1}$ with $||u|| = 1$. Then instead of solving for $\frac{\partial}{\partial\beta}\sum_{i=1}^n \rho_\tau(Y_i - x_i'\beta) = 0$ the minimization problem can be restated as finding $\hat{\beta}$ such that

$$\frac{\partial}{\partial a}\rho_\tau(Y_i - x_i'\hat{\beta} - ax_i'w)\Big|_{a=0} \geq 0 \ \forall w. \tag{2.5}$$

Let $Q(\beta) = \sum_{i=1}^n \rho_\tau(\beta)$. If $\hat{\beta}$ satisfies (2.5) then $Q(\hat{\beta})$ is a local minimum and because $Q(\beta)$ is a convex function $Q(\hat{\beta})$ is also a global minimum. The solution space can be limited to cases where $p+1$ observations, corresponding to the $p+1$ parameters being estimated, have residuals of zero. Let Ω be the set of the different combinations of $p+1$ observations from a sample of size n. Let $\omega \in \Omega$ represent one such combination and $X(\omega)$ and $Y(\omega)$ be the corresponding covariates and response. Define

$$\beta(\omega) = X(\omega)^{-1}Y(\omega).$$

Therefore $X(\omega)\beta(\omega) = Y(\omega)$ and $\beta(\omega)$ is a candidate for $\hat{\beta}$. Let $\mathcal{B} = \{\beta(\omega) \mid \omega \in \Omega\}$. Then (2.3) could be restated as

$$\hat{\beta} = \underset{\beta\in\mathcal{B}}{\operatorname{argmin}} \sum_{i=1}^n \rho_\tau(Y_i - x_i'\beta).$$

The number of potential solutions has been reduced from ∞ to $\binom{n}{p+1}$, but checking every $\beta(\omega) \in \mathcal{B}$ would not be practical for larger sample sizes.(Koenker, 2005) Koenker and d'Orey (1987) presented a modified algorithm of Barrodale and Roberts (1974), which proposed an algorithm for median regression. First start with an initial estimate

$\beta(\omega)$ and evaluate the partial derivative of $Q(\beta(\omega))$. Next find the path of steepest descent. Since $Q(\beta(\omega))$ is a vertex of a convex function directions can be limited to the edges that meet at the vertex. The edge of quickest decent is followed until it is no longer a viable path, thus arriving at another vertex where the algorithm can be repeated. The algorithm stops once it hits a vertex where all edge, directional derivatives are positive. This algorithm was critical to the development of median and quantile regression because it provided an efficient method for estimating regression quantiles.

To visualize the algorithm we consider a simple example in the unconditional setting. Take two samples, one with nine observations of $\{1, 3, 5, 7, 9, 13, 15, 17, 19\}$ and the other with eight observations of $\{1, 3, 5, 7, 13, 15, 17, 19\}$. The quantile regression objective function for an intercept only model was applied to both data sets for the median and .25 quantile. Figure 2.5 has plots of the objective function for the four different scenarios. Those marked as having unique solutions are from the first sample, while the non-unique solutions come from the second sample. The function is clearly not differentiable at the observed values in the sample set, which would be the set of potential solutions. The non-unique set shows that there are solutions that would not have a residual of zero, but these solutions all lie on an edge between two vertexes. Using the algorithm from Koenker and d'Orey (1987) results in different solutions for the non-unique case depending on the starting point. For the median case with the second sample $\hat{\beta} = 7$ if the initial value is 7 or less, $\hat{\beta} = 13$ if the initial value is 13 or larger and $\hat{\beta} =$ initial value for initial values between 7 and 13. The issue of a non-unique solution also occurs when minimizing a conditional objective function. Koenker (2012) argued that the flat edge of non-unique solutions are small compared to sampling error from the data because the flat edge of non-unique solutions becomes smaller as the sample size increases.

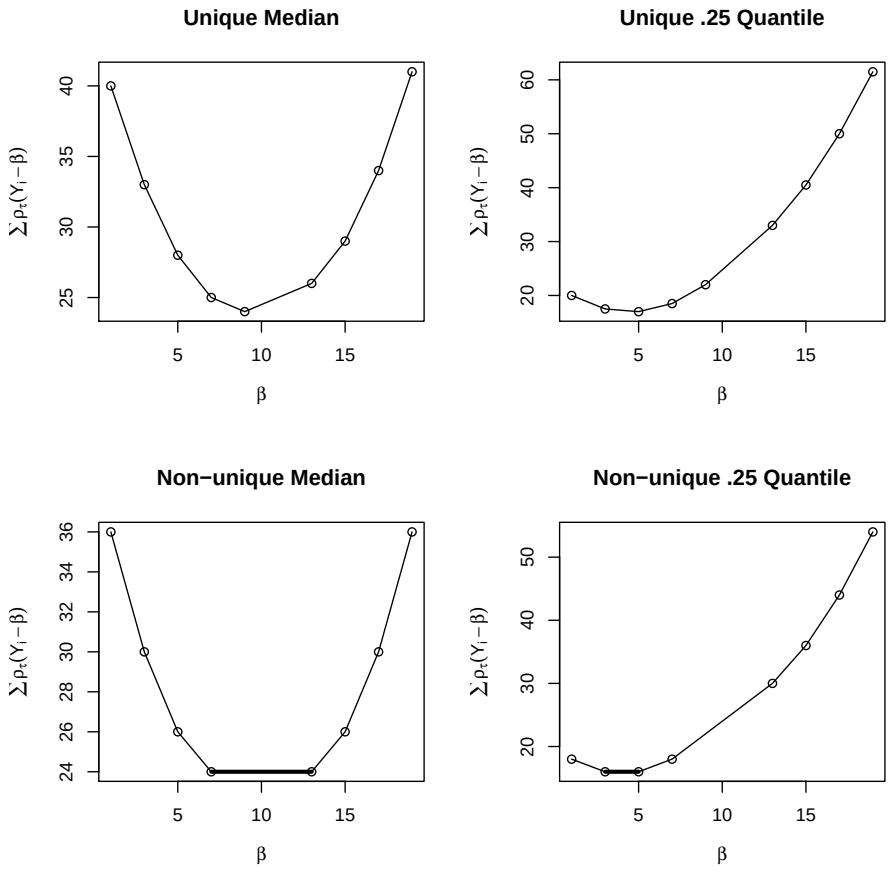

Figure 2.5: Unconditional Minimization

2.4.2 Theoretical Challenges

Not being able to take a derivative also provides theoretical difficulties. A simple proof of the asymptotic normality of $\hat{\beta}(\mu)$ relies on being able to take a derivative of the objective function.

$$\begin{aligned} & -n^{-1}\textstyle\sum_{i=1}^{n} x_i(Y_i - x_i'\hat{\beta}(\mu)) = 0. \\ \Rightarrow \quad & -n^{-1}\textstyle\sum_{i=1}^{n} x_i(\epsilon_i + x_i'\beta_0 - x_i'\hat{\beta}(\mu)) = 0. \\ \Rightarrow \quad & n^{-1}\textstyle\sum_{i=1}^{n} x_i x_i'(\hat{\beta}(\mu) - \beta_0) = n^{-1}\textstyle\sum_{i=1}^{n} x_i\epsilon_i. \\ \Rightarrow \quad & (n^{-1}X'X)\sqrt{n}(\hat{\beta}(\mu) - \beta_0) = n^{-1/2}\textstyle\sum_{i=1}^{n} x_i\epsilon_i. \end{aligned}$$

Then assuming $\frac{1}{n}X'X \xrightarrow{p} \Sigma$ and $E[\epsilon_i] = \sigma^2$

$$\sqrt{n}(\hat{\beta}(\mu) - \beta_0) \xrightarrow{d} N\left(0, \sigma^2\Sigma^{-1}\right).$$

Deriving theoretical properties of quantile regression estimators requires more subtle methods which typically rely on convexity of the objective function. The proof of Theorem 4.1 of Koenker (2005) outlines an approach to analyze the asymptotic behavior of estimator from a convex objective function. The central idea is to approximate the objective function with a quadratic function. Hjørt and Pollard (1993) showed that if a convex function can be approximated by a quadratic function the minimizer of the quadratic function is asymptotically equivalent to the minimizer of the convex function. Thus reducing the problem to an easier problem of understanding the asymptotic behavior of the minimizer of a quadratic approximation.

For technical reasons it is helpful to restate (2.3) as

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} \rho_\tau(Y_i - x_i'\beta) - \rho_\tau(Y_i - x_i'\beta_0). \tag{2.6}$$

Let F_i and f_i be the CDF and pdf of $\epsilon_i \mid x_i$. Notice that $f_i(0)$ is the density at the τth quantile of interest. Assume that F_i is absolutely continuous, f_i is uniformly bounded away from 0 and ∞ and $\forall i$ $x_i \in D$ where D is a compact subspace of $\mathcal{R}^{p+1}$. Also define $\psi_\tau(u) = \tau - I(u < 0)$, the gradient of $\rho_\tau(u)$. Then using Knight's Identity (Knight, 1998),

$$\rho_\tau(u - v) - \rho_\tau(u) = -v\psi_\tau(u) + \int_0^v (I(u \leq s) - I(u \leq 0))ds,$$

and Taylor expansion we have

$$\sum_{i=1}^n \rho_\tau(Y_i - x_i'\beta) - \rho_\tau(Y_i - x_i'\beta_0) = \frac{1}{2}(\beta - \beta_0)' \sum_{i=1}^n f_i(0)x_i x_i'(\beta - \beta_0) - (\beta - \beta_0)' \sum_{i=1}^n x_i\psi_\tau(\epsilon_i) + o_p(1). \tag{2.7}$$

Then behavior of $\hat{\beta}$ from (2.3) is asymptotically equivalent to the minimizer of (2.7) thus

$$\left[\frac{1}{n}\sum_{i=1}^n f_i(0)x_i x_i'\right]\sqrt{n}(\hat{\beta} - \beta_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^n x_i\psi_\tau(\epsilon_i) + o_p(1). \tag{2.8}$$

If $\frac{1}{n}\sum_{i=1}^n f_i(0)x_i x_i' \xrightarrow{p} \tilde{\Sigma}$ and $\frac{1}{n}\sum_{i=1}^n x_i x_i' \xrightarrow{p} \Sigma$ then

$$\sqrt{n}\left(\hat{\beta} - \beta_0\right) \xrightarrow{d} N\left(0, \tau(1-\tau)\tilde{\Sigma}^{-1}\Sigma\tilde{\Sigma}^{-1}\right).$$

The $\tau(1-\tau)$ portion of the asymptotic confirms the intuition that the estimates further from $\tau = 1/2$ will tend to have larger variance. If we assume that ϵ_i are $i.i.d.$ then

$$\sqrt{n}\left(\hat{\beta} - \beta_0\right) \xrightarrow{d} N\left(0, \tau(1-\tau)f_i(0)^{-2}\Sigma^{-1}\right).$$

Recall, that $f_i(0)$ is the density of $\epsilon_i \mid x_i$ at the conditional quantile of interest. The estimator $\hat{\beta}$ has larger variance when estimating events that have small density. This typically happens when modeling higher or lower quantiles.

Chapter 3

Additive Partial Linear Quantile Regression

In an additive partial linear regression model, there is a set of predictors and constant $X \in \mathbb{R}^{p+1}$ which have a linear relationship with the response variable $Y \in \mathbb{R}$ and a set of predictors $Z \in \mathbb{R}^d$ which have an unknown non-linear relationship with Y, described by the nonparametric component $g(Z)$. Formally, the additive partial linear quantile regression model is

$$\begin{aligned} Y_i &= \beta_{00}(\tau) + \beta_{10}(\tau)x_{i1} + ...\beta_{p0}(\tau)x_{ip} + g_0(\tau, z_i) + \epsilon_i \\ &= x_i'\beta_0(\tau) + g_0(\tau, z_i) + \epsilon_i, \end{aligned}$$

with $x_i = (1, x_{i1}, ..., x_{ip})'$, $z_i = (z_{i1}, ..., z_{id})'$, $\beta_0(\tau) = (\beta_{00}(\tau), \beta_{10}(\tau), ..., \beta_{p0}(\tau))'$ and $P(\epsilon_i \leq 0 | x_i, z_i) = \tau$, for some $0 < \tau < 1$. To avoid the "curse of dimensionality" we assume that $g_0(\tau, z_i)$ is an additive function where

$$g_0(\tau, z_i) = \sum_{j=1}^{d} g_{j0}(\tau, z_{ij}).$$

The additive partial linear model balances the flexibility of non-parametric methods with the ease of interpretation of parametric models. One application of this model is

to include variables that require easy interpretation as linear variables and nuisance variables as non-linear. For example, an experiment with a binary treatment and continuous controls. The binary treatment would be treated as a linear variable. The control variables could be modeled as unknown relationships because we are more concerned about bias from a misspecified model for these terms and less concerned about interpretation.

3.1 Basis Splines

Each function $g_{j0}(\tau, z)$ is unknown and needs to be estimated. We consider estimation using series methods which approximates $g_{j0}(\tau, z)$ by a linear combination of a series of J_n, where J_n can change with n, approximating functions $p_i(z)$, $i = 1, \ldots, J_n$. Polynomial functions are popular choices for the approximating functions. A simple and classic example of a series estimate is using the power series $\{1, z, z^2, ..., z^{J_n-1}\}$. A problem with the power series it that successive terms tend to be highly correlated. We focus on using B-splines which are a linear combination of a set of basis splines. B-spline functions are piecewise polynomial functions and generally provide more flexibility than the power series. They are also numerically more stable because each B-spline is non-zero over a limited range of knots. To define the B-spline functions, we divide the observed support of z into m_n intervals and let r be the degree of the functions used. Let $(t_1, ..., t_{2r+m-1})$ be our sequence of knots with $m_n - 1$ knots inside the compact support and r knots on the lower bound and upper bound of the support, for a total of $J_n = r + m_n$ basis functions.(Schumaker, 1981) The formula for basis

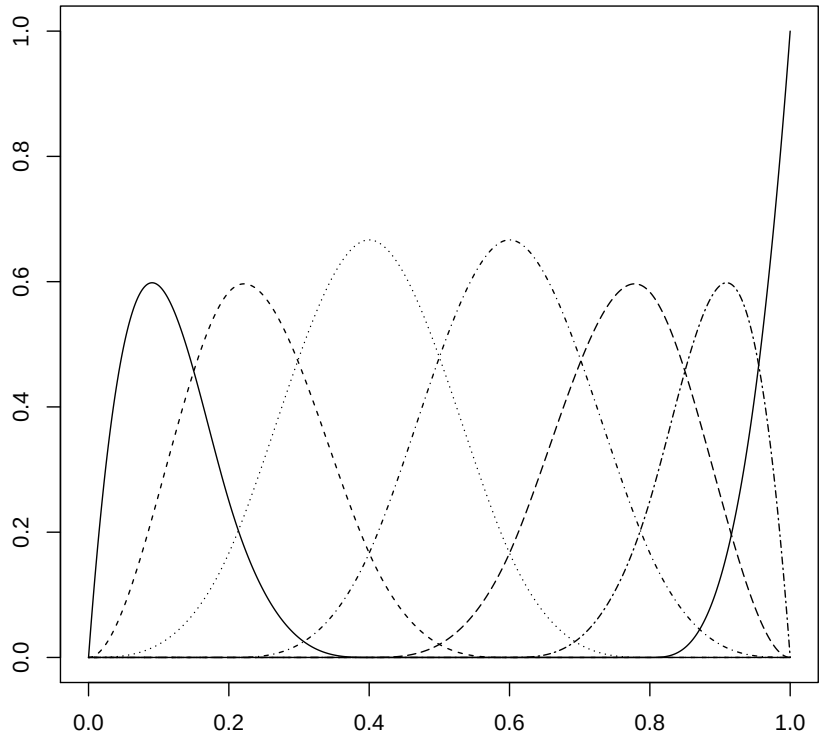

Figure 3.1: Plot of Cubic Splines with $J_n = 7$

functions are defined by the following recursive formula:

$$
\begin{aligned}
b_1^r(z) &= \begin{cases} 1 & t_i \leq z \leq t_{i+1}, \\ 0 & \text{otherwise}, \end{cases} \\
b_i^r(z) &= \frac{z - t_i}{t_{i+r-1} - t_i} b_i^{r-1}(z) + \frac{t_{i+r} - z}{t_{i+r} - t_{i+1}} b_{i+1}^{r-1}(z).
\end{aligned}
$$

Figure 3.1 displays seven evenly spaced cubic B-splines on a support of $[0, 1]$.

For a given covariate z_{ik} and degree r let $w(z_{ik}) = \left(b_1^r(z_{ik}), ..., b_{J_n}^r(z_{ik})\right)'$ denote the corresponding vector of B-spline basis functions and let $w(z_i)$ denote the dJ_n

dimensional vector $(w(z_{i1})', ..., w(z_{id})')'$. For ease of notation and simplicity of proofs, we use the same basis functions and J_n for all non-linear components. Then (3.1) is estimated by

$$(\hat{\beta}, \hat{\gamma}) = \underset{(\beta,\gamma)}{\operatorname{argmin}} \sum_{i=1}^{n} \rho_\tau(Y_i - x_i'\beta - w(z_i)'\gamma) \tag{3.1}$$

with

$$\hat{g}_j(z_i) = w_j(z_i)'\hat{\gamma}_j - \frac{1}{n}\sum_{k=1}^{n} w_j(z_k)'\hat{\gamma}_j \text{ for } j = 1, ..., d,$$

where $\hat{\gamma}_j$ is the basis coefficients corresponding to $w_j(z_i)$. The estimate of $\hat{g}$ is centered because of the identifiability condition that $E[g_j(z_i)] = 0 \, \forall j$. The intercept needs to be adjusted by $\frac{1}{n}\sum_{i=1}^{n} w(z_i)'\hat{\gamma}$ to account for the centering. To avoid these complications when dealing with the asymptotic behavior of these estimators we use centered and standardized B-splines, following the approach of Liu et al. (2011). The centered spline for the jth basis function of the covariate z_{ik} is

$$b_j^*(z_{ik}) = b_j(z_{ik})^r - \frac{E\left[b_j^r(z_{ik})\right]}{E\left[b_1^r(z_{ik})\right]} b_1^r(z_{ik}), \tag{3.2}$$

suppressing the degree r for $b_j^*(z_{ik})$ for ease of notation. The centered and standardized spline is

$$B_j(z_{ik}) = \frac{b_j^*(z_{ik})}{\sqrt{\text{Var}(b_j^*(z_{ik}))}}. \tag{3.3}$$

Let $W(z_{ik}) = (B_1(z_{ik}), ..., B_{J_n}(z_{ik}))'$ denote the corresponding vector of centered and standardized B-spline basis functions and define $W(z_i)$ as the dJ_n dimensional vector

$(W(z_{i1})', ..., W(z_{id})')'$. Then the estimators $\hat{g}$ and $\hat{\beta}$ obtained from minimizing

$$\underset{(\beta,\gamma)}{\operatorname{argmin}} \sum_{i=1}^{n} \rho_\tau(Y_i - x_i'\beta - W(z_i)'\gamma), \tag{3.4}$$

are the same as those from minimizing (3.1) and centering $\hat{g}$ and the intercept. In practice we obtain our estimates from (3.1) because the value of $B_j(z_{ik})$ is unknown. However for theoretical reasons it will be easier to use the equivalent estimators derived from (3.4). For this reason we will use $w(z_i)$ when referring to uses of the additive partial linear quantile regression model in practice, but in our proofs we will use $W(z_i)$.

A complication of the partial-linear model is the estimation error for the non-linear component. For a fixed n there is an idealized $\gamma_0 \in \mathbb{R}^{dJ_n}$, but in general $g_0(z_i) \neq W(z_i)'\gamma_0$ and instead

$$W(z_i)'\gamma_0 - g_0(z_i) = u_{ni}, \tag{3.5}$$

where u_{ni} is the bias term. To understand the behavior of u_{ni} we require the two following definitions.

Definition Let $r \equiv m + v$. Define $\mathcal{H}_r$ as the collection of functions on $[0, 1]$ whose mth derivative satisfy the Hölder condition of order v. That is, for any $h \in \mathcal{H}_r$, there exist some positive constant C such that

$$\left|h^{(m)}(z') - h^{(m)}(z)\right| \leq C\left|z' - z\right|^v, \quad \forall \quad 0 \leq z', z \leq 1. \tag{3.6}$$

Definition Given $z = (z_1, ..., z_d)'$, the function $g(z)$ is said to belong to the class of non-linear functions $\mathcal{G}_r$ if $g(z) = \sum_{k=1}^{d} g_k(z_k)$, $g_k \in \mathcal{H}_r$ and $E[g_k(z_k)] = 0 \ \forall \ k$.

Throughout this chapter we assume $g_0 \in \mathcal{G}_r$ for some $r \geq 1.5$. Then the function g_0

can be approximated using B-spline basis functions and the bias term has a rate of convergence of $\max_i |u_{ni}| = O\left(J_n^{-r}\right)$.(Schumaker, 1981) Nonparametric mean models have been an active area of research. Consider the univariate, $z_i \in \mathbb{R}$, mean model

$$Y_i = g_0(z_i) + \epsilon_i, \tag{3.7}$$

with $E[\epsilon_i] = 0$. Stone (1982) showed that for $J_n \approx n^{1/(2r+1)}$ that

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{g}(z_i) - g_0(z_i))^2 = O_p\left(n^{-2r/(2r+1)}\right), \tag{3.8}$$

and that (3.8) is the optimal rate of convergence. The intuition is that the rate of convergence of $\frac{1}{n}\sum_{i=1}^{n}(\hat{g}(z_i) - g_0(z_i))^2$ can be separated into the estimation of the spline coefficient vector and the rate of convergence for the bias term, u_{ni}. The spline coefficient vector has the rate of $||\hat{\gamma} - \gamma_0|| = O_p\left(\sqrt{\frac{J_n}{n}}\right)$. The rate of J_n which minimizes both rates is $J_n \approx n^{1/(2r+1)}$ which combined with the rate of the spline coefficients and bias provides the rate given in (3.8). It has also been shown that for an additive version of (3.7) that the rate of (3.8) holds.(Stone, 1985)

Donald and Newey (1994) proposed a partial linear mean model and did not assume the non-linear function was additive. This work showed conditions for $\hat{\beta}$ to be asymptotically normal and efficient estimation of g_0. He and Shi (1994) demonstrated that (3.8) holds for non-parametric estimates of a univariate conditional quantile function. These results were extended to partial linear quantile regression (He and Shi, 1996) and partial linear m-estimation models (He et al., 2002). Wang et al. (2009) consider a partial linear varying coefficient model and propose a penalized procedure for variable selection of the linear terms. De Gooijer and Zerom (2003) developed a fully non-parametric additive quantile regression estimator that has the same asymptotic rate of convergence as the univariate estimator proposed by He

and Shi (1994). However the method requires bias correction for $d \geq 5$. Alternative estimators for non-parametric additive quantile regression models have been proposed that retain efficient estimation and do not have to correct for bias.(Horowitz and Lee, 2005) In the current literature there have not been any asymptotic results for additive partial linear quantile regression models. We demonstrate that under standard regularity conditions the estimators from (3.1) are consistent and $\hat{\beta}$ is asymptotically normal. To understand the asymptotic behavior of $\hat{\beta}$ we need to establish a relationship between X and Z.

3.2 Relationship between X and Z

Before considering the assumptions that are needed for the additive partial linear quantile regression model we start with the additive mean regression model to motivate these assumptions. Consider the least squares objective function of

$$\left(\hat{\beta}(\mu), \hat{\gamma}(\mu)\right) = \underset{(\beta,\gamma)}{\operatorname{argmin}} \sum_{i=1}^{n} (Y_i - x_i'\beta - W(z_i)'\gamma)^2. \tag{3.9}$$

To understand the asymptotic behavior of $\hat{\beta}(\mu)$ we need to consider the relationship between X and Z. New notation is introduced to separate the constant part of X from the random portion. Let $X = [1_n \; X_{(-1)}]$ where 1_n is an n-dimensional vector of ones and $X_{(-1)} \in \mathbb{R}^{n \times p}$ with $X_{(-1)} = (X_1, ..., X_p)$. Let

$$x_{ij} = h_{j0}^{\mu}(z_i) + \delta_{ij}^{\mu} \; 1 \leq i \leq n, \; 1 \leq j \leq p,$$

with $h_{j0}^{\mu} \in \mathcal{H}_r^d$ and δ_{ij}^{μ} being the bias from estimating $E[x_{ij} \mid z_i]$ with an additive function of z_i. To handle the intercept define $h_{10}^{\mu}(z_i) = 0$ and $\delta_{i1}^{\mu} = 1 \; \forall i$. Let

$H(\mu)ij = h^{\mu}_{(j+1)0}(z_i)$, $\delta^{\mu}_i = \left(1, \delta^{\mu}_{i2}, ..., \delta^{\mu}_{i(p+1)}\right)'$ and $\Delta(\mu)_{ij} = (\delta^{\mu}_1, ..., \delta^{\mu}_n)'$ then

$$X = H(\mu) + \Delta(\mu).$$

Define $W = (W(z_1), ..., W(z_n))'$ and

$$\begin{aligned} P_W &= W'(W'W)^{-1}W, \\ X^*(\mu) &= [1_n, (I - P_W)X_{(-1)}], \end{aligned}$$

with $X^*(\mu) = (x^*_1(\mu)', ..., x^*_n(\mu)')'$. Consider the follow parametrization

$$\begin{aligned} \theta_1(\mu) &= (\beta - \beta_0(\mu)), \\ \theta_2(\mu) &= (W'W)^{-1}WX(\beta - \beta_0(\mu)) + (\gamma - \gamma_0(\mu)), \end{aligned}$$

with

$$\hat{\theta}_1(\mu) = (\hat{\beta}(\mu) - \beta_0(\mu))$$

and

$$\hat{\theta}_2(\mu) = (W'W)^{-1}WX(\hat{\beta}(\mu) - \beta_0(\mu)) + (\hat{\gamma}(\mu) - \gamma_0(\mu)).$$

Then (3.9) is equivalent to

$$\left(\hat{\theta}_1(\mu), \hat{\theta}_2(\mu)\right) = \underset{(\theta_1,\theta_2)}{\operatorname{argmin}} \sum_{i=1}^{n} (\epsilon_i - x^*_i(\mu)'\theta_1 - W(z_i)'\theta_2 - u_{ni})^2 \tag{3.10}$$

In order to find the asymptotic behavior of $\hat{\beta}$ we can solve (3.10) with respect to θ_1 and get

$$\sum_{i=1}^{n} x_i^*(\mu)(\epsilon_i - x_i^*(\mu)'\hat{\theta}_1(\mu) - W(z_i)'\hat{\theta}_2 - u_{ni})) = 0.$$

Notice that $\sum_{i=1}^{n} x_i^*(\mu)W(z_i)' = X^*(\mu)'W = X'(I - P_W)W = 0$ and therefore

$$\left(\frac{1}{n}\sum_{i=1}^{n} x_i^*(\mu)x_i^*(\mu)'\right)\sqrt{n}\hat{\theta}_1(\mu) = n^{-1/2}\sum_{i=1}^{n} x_i^*(\mu)\epsilon_i - n^{-1/2}\sum_{i=1}^{n} x_i^*(\mu)u_{ni}.$$

Thus $\sqrt{n}(\hat{\beta}(\mu) - \beta_0(\mu))$ is asymptotically normal if $\max_i ||x_i^*(\mu)u_{ni}|| = o_p(n^{-1/2})$ and $\frac{1}{n}\sum_{i=1}^{n} x_i^*(\mu)x_i^*(\mu)'$ converges in probability to a positive definite matrix. The former can be established by reasonable assumptions for X and selecting J_n at a suitable rate. Recall $H(\mu)$ is an additive approximation of $E[X \mid Z]$ and $P_W X$ is the least squares estimate of $H(\mu)$. Therefore under conditions described in Stone (1985) $n^{-1/2}x_i^*(\mu) = n^{-1/2}\delta_i^{\mu} + o_p(1)$.

Using a similar parametrization and applying methods used to derive (2.7) then

$$\begin{aligned}
&\sum_{i=1}^{n} \rho_\tau(\epsilon_i - x_i^{*\prime}\theta_1 - W(z_i)'\theta_2 - u_{ni}) - \rho_\tau(\epsilon_i) \\
= &\sum_{i=1}^{n} \left(x_i^{*\prime}\theta_1 + W(z_i)'\theta_2 - u_{ni}\right)\psi_\tau(\epsilon_i) \\
+ &\sum_{i=1}^{n} f_i(0)\left(x_i^{*\prime}\theta_1 + W(z_i)'\theta_2 - u_{ni}\right)^2 + o_p(1).
\end{aligned}$$

Similar to the least squares case we want to solve for $\hat{\theta}_1$ and use this solution to derive asymptotic normality for $\hat{\beta}$. Understanding the asymptotic behavior of $\hat{\theta}_1$ is easier if

$$\sum_{i=1}^{n} f_i(0)x_i^*W(z_i) = 0. \tag{3.11}$$

Define $B = \text{diag}(f_1(0), ..., f_n(0))$ and $P_W(B) = W(W'BW)^{-1}W'B$. Then (3.11) holds if $X^* = \left[1_n \ (I - P_W(B))X_{(-1)}\right]$. Therefore using this technique for an additive partial linear quantile regression model requires a different understanding of the role estimating g_0 has on the asymptotic behavior of $\hat{\beta}$. Let $X = H + \Delta_n$ with $H_{ij} = h_{j+1}(z_i)$, $\delta_i = (1, \delta_{i1}, ..., \delta_{ip})'$, $\Delta_n = (\delta_1, ..., \delta_n)'$ and

$$\begin{aligned} h_j(\cdot) &= \arg\inf_{h_j \in \mathcal{H}_r^d} \sum_{i=1}^{n} E\left[f_i(0)(x_{ij} - h_j(z_i))^2\right] \ \ 1 \leq j \leq p, \\ h_0(\cdot) &= 0. \end{aligned}$$

Then $P_W(B)X$ is a weighted least squares estimate of H and applying the results of Stone (1985) $n^{-1/2}x_i^* = n^{-1/2}\delta_i + o_p(1)$. Thus creating a similar setup for quantile regression which allows $f_i(0)$ to be non-constant.

3.3 Conditions

The following conditions are assumed to understand the behavior of $\hat{\beta}$ and $\hat{g}$.

Condition 1

(Conditions on the random error) The random error ϵ_i has distribution function F_i and continuous density function f_i. The f_i are uniformly bounded away from 0 and infinity in a neighborhood of zero, its first derivative f_i' has a uniform upper bound in a neighborhood of zero, for $1 \leq i \leq n$. □

Condition 2

(Conditions on the covariates) There exist a positive constant M_1 such that $|x_{ij}| \leq M_1$, $\forall\ 1 \leq i \leq n$, $1 \leq j \leq p$. □

Condition 3

(Condition on the non-linear functions) For $r = m + v > 1.5$ $g_0 \in \mathcal{G}_r$ and $\forall\ j$, $h_j \in \mathcal{H}_r^d$. □

Condition 4

(Condition on the B-spline basis) The dimension of the spline basis J_n has the following rate

$$n^{1/2r} \ll J_n \ll n^{1/3}. \qquad \square$$

Condition 5

(Condition for asymptotic covariance) For positive definite matrices Σ_1 and Σ_2

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n} f_i(0)\delta_i\delta_i' &\xrightarrow{p} \Sigma_1, \\ \frac{1}{n}\tau(1-\tau)\sum_{i=1}^{n} \delta_i\delta_i' &\xrightarrow{p} \Sigma_2. \end{aligned}$$

Conditions 1 and 2 are common quantile regression assumptions. Condition 3 allows results from Stone (1985) to be used to for estimating g and for theoretical reasons h. Condition 4 results in an undersmoothed estimate of g which is convenient for proving that $\hat{\beta}$ is asymptotically normal. Condition 5 is needed to define the asymptotic covariance of $\hat{\beta}$.

3.4 Asymptotic Results

The following theorems summarize the asymptotic properties of the estimators from (3.1).

Theorem 3.1

If conditions 1-5 hold then for the estimators from (3.1)

$$\begin{aligned} ||\hat{\beta} - \beta_0|| &= o_p(1), \\ \frac{1}{n}\sum_{i=1}^{n}(\hat{g}(z_i) - g_0(z_i))^2 &= O_p\left(\frac{J_n}{n}\right). \end{aligned}$$

Our proof of Theorem 3.1 allows for $J_n \approx n^{1/(2r+1)}$ which provides the optimal rate of convergence for $\hat{g}$. However, our proof of asymptotic normality of $\hat{\beta}$ requires condition 4 which does not allow for $J_n \approx n^{1/(2r+1)}$.

Theorem 3.2

If conditions 1-5 hold then for $\hat{\beta}$ from (3.1)

$$\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{d} N(0, \Sigma_1^{-1}\Sigma_2\Sigma_1^{-1}). \qquad \square$$

He et al. (2002) contains similar results to Theorem 3.1 and Theorem 3.2 for a partial linear longitudinal model, but only considers $d = 1$. In Chapter 6 we discuss extending this model to a longitudinal setting and other future research directions.

3.5 Simulations

We tested the additive partial linear model under a variety of simulation settings. Quantile regression models are fit for $\tau = .5$ and $\tau = .7$ and are compared to mean regression fits. For the simulations we generate $X_1 \sim \text{Ber}(.5)$, $X_2 \sim N(0,1)$, $X_3 \sim N(0,1)$, $Z_1 \sim U[0,1]$ and $Z_2 \sim U[-1,1]$. For $i = 1, ..., n$ the response is generated by

$$Y_i = 3x_{i1} + 1.5x_{i2} + 2x_{i5} + sin(2\pi z_{i1}) + z_{i2}^3 + \epsilon_i.$$

We consider three different distributions for ϵ_i: (1) standard normal distribution; (2) t distribution with three degrees of freedom; (3) heteroscedastic normal distribution with $\epsilon_i = (1 + x_{i1})\xi_i$ where $\xi_i \sim N(0, 1)$ are independent of the x_i's. These three cases allow us to evaluate our estimators performance for the most favorable setting for mean regression, for a heavy-tailed distribution and a case with heteroscedastic errors.

The following criteria are used to assess the performance of the estimators:

1. AADE: average of the *average absolute deviation* (ADE) of the fit of the non-linear components, where the ADE is defined as $n^{-1}\sum_{i=1}^{n}|\hat{g}(z_i) - g_0(z_i)|$.

2. MSE: average of the mean squared error for estimating β_0, that is, average of $||\hat{\beta} - \beta_0||^2$.

3. $\hat{\beta}_1$: Average of $\hat{\beta}_1$, the estimate for the coefficient of x_1.

The three different methods used are

1. $Q_{.5}$: Quantile regression for the median,

2. $Q_{.7}$: Quantile regression for $\tau = .7$,

3. MR: Mean regression using OLS.

In all simulations coefficients were the same for median and mean regression. For the heteroscedastic setting the value of β_{10}, the coefficient of x_1, changes with τ. Therefore $\hat{\beta}_1$ is reported for the heteroscedastic case because quantile regression can identify that the coefficients of x_1 changes with τ, while OLS can not. Simulations were run for samples sizes of $n = 100, 300$, or 1000 and 100 simulations were run for each setting. In each simulation we considered 3 to 15 basis functions for Z_1 and Z_2. Let J_{1_n} and J_{2_n} be the number of basis functions used for Z_1 and Z_2. Let $\hat{\beta}(J_{1_n}, J_{2_n})$ and $\hat{\gamma}(J_{1_n}, J_{2_n})$ be the estimates derived when using J_{1_n} and J_{2_n} basis

functions. Further let $\nu(J_{1_n}, J_{2_n}) = 3 + J_{1_n} + J_{2_n}$, the degrees of freedom in a model with J_{n_1} and J_{n_2}. Define

$$\begin{aligned} QBIC(J_{1_n}, J_{2_n}) &= \log\left(\sum_{i=1}^{n} \rho_\tau\left(Y_i - x_i'\hat{\beta}(J_{1_n}, J_{2_n}) - w(z_i)'\hat{\gamma}(J_{1_n}, J_{2_n})\right)\right) \\ &+ \frac{\nu(J_{1_n}, J_{2_n})\log(n)}{2n}. \end{aligned}$$

The final model is selected by finding the combination of J_{1_n} and J_{2_n} that minimizes $QBIC(\cdot)$. Horowitz and Lee (2005) proposed a similar BIC type method for a fully non-parametric additive quantile regression model.

Results of the simulations are reported in Table 3.1 - Table 3.3. In all simulations we see that estimation accuracy increases with n and that estimates for $\tau = .5$ are more accurate then those for $\tau = .7$. For the case of $\epsilon_i \sim N(0, 1)$ mean regression outperforms quantile regression. For $\epsilon_i \sim T_3$ median regression outperforms mean regression. Quantile regression for $\tau = .7$ has similar efficiency for estimates of β, but mean regression does outperform it for estimation of the nonlinear functions. For the case of the heteroscedastic error $\beta_{10}(.7) \approx 3.52$, the coefficient for x_1 when $\tau = .7$. Table 3.3 shows that $\hat{\beta}_1(.7)$ provides a consistent estimate of this value but the least squares method is biased. Focusing only on mean regression loses the nuance that some of the coefficients change with τ because of heteroscedasticity.

Method	n	AADE	MSE
$Q_{.5}$	100	0.29	0.09
$Q_{.5}$	300	0.18	0.03
$Q_{.5}$	1000	0.11	0.01
$Q_{.7}$	100	0.54	0.11
$Q_{.7}$	300	0.52	0.03
$Q_{.7}$	1000	0.51	0.01
MR	100	0.24	0.07
MR	300	0.14	0.02
MR	1000	0.09	0.01

Table 3.1: Additive Partial Linear Simulation Results for $\epsilon_i \sim N(0,1)$

Method	n	AADE	MSE
$Q_{.5}$	100	0.33	0.14
$Q_{.5}$	300	0.18	0.03
$Q_{.5}$	1000	0.11	0.01
$Q_{.7}$	100	0.66	0.2
$Q_{.7}$	300	0.58	0.05
$Q_{.7}$	1000	0.57	0.02
MR	100	0.37	0.19
MR	300	0.21	0.07
MR	1000	0.13	0.02

Table 3.2: Additive Partial Linear Simulation Results for $\epsilon_i \sim T_3$

Method	n	$\hat{\beta}_1$	AADE	MSE
$Q_{.5}$	100	2.99	0.37	0.24
$Q_{.5}$	300	3.02	0.21	0.06
$Q_{.5}$	1000	3.00	0.12	0.02
$Q_{.7}$	100	3.45	0.6	0.27
$Q_{.7}$	300	3.56	0.52	0.07
$Q_{.7}$	1000	3.54	0.51	0.02
MR	100	3.01	0.31	0.18
MR	300	3.00	0.19	0.04
MR	1000	2.99	0.12	0.02

Table 3.3: Additive Partial Linear Simulation Results for Heteroscedastic ϵ_i

3.6 Proofs

Theorem 3.1 and Theorem 3.2 are special cases of Theorem 5.1 and Theorem 5.2 and results follow from proofs provided in Chapter 5.

Chapter 4

Quantile Regression with Missing Covariates

Missing data is a common problem in data analysis. If subjects with any missing values are dropped from the analysis this will lead to a biased analysis if there is a systematic pattern to the missingness. We focus on the case of missing covariates, which can occur for a variety of reasons. For example in a medical study people may refuse to answer certain questions, a nurse may forget to make all of the measurements or subjects may miss a follow-up appointment.

Two popular methods for handling missing data are weighting and imputation. The imputation approach replaces the missing values with imputed values and performs the analysis as if the data were complete. Weighting methods reweight the records with complete data to account for the bias from ignoring records with missing data. The imputation approach is often based on likelihood analysis and requires specifying a joint or conditional likelihood. Although the likelihood-based imputation method is usually more efficient than the weighting method, correct specification of the joint likelihood function is often challenging in practice. This is particularly a problem for skewed and heteroscedastic data, a setting where quantile regression is particularly useful. When the likelihood function is misspecified, the imputation approach may lead to biased estimation. The quantile regression based weighting

approach we propose is semiparametric and circumvents the difficulty of specifying the joint likelihood function. In particular, it requires no parametric assumptions for the covariates or the error term. The main idea is *inverse probability weighting* (IPW), that is, we weight the completely observed cases inversely proportionally to the probability of being observed. Existing work has demonstrated that linear mean regression estimator using IPW is asymptotically normal.(Robins et al., 1994) The method can also be extended to semiparametric and nonparametric mean regression models.(Tsiatis, 2006)

Research on how to handle missing data when using quantile regression is limited. A multiple imputation method has been proposed which alleviates the decrease in efficiency caused by missing data. However, the method assumes the missing is completely random and thus does not deal with the issue of bias caused by missing data.(Wei et al., 2012) Lipsitz et al. (1997) and Yi and He (2009) studied IPW methods for longitudinal quantile regression models with dropouts where the covariates are time invariant, thus are known at all time points, but the response variable may be missing from a certain time point. The weighted estimators proposed in these two papers are defined by weighted estimating equations. We consider the case of covariates missing at random and study an estimator defined as the minimizer of a weighted quantile objective function. In our earlier work we proposed a weighted method for the linear model assuming a logistic regression model for the missing model and proposed a modified BIC for variable selection in the presence of missing data.(Sherwood et al., 2013) In this chapter we analyze the more flexible additive partial linear model and relax the assumption that a logistic regression model is needed to model the rate of missingness. For model selection we consider an objective function with penalties for the linear terms. Liang et al. (2004) proposed a penalized objective function for model selection of the linear terms for a partial linear mean regression. Wu and Liu (2009) proposed using a penalized objective function for

variable selection in quantile regression. The current literature does not cover model selection for an additive partial linear quantile regression nor the use of penalized objective functions for missing data problems.

4.1 Bias from Missing Data

Statisticians often consider three types of missingness: missing completely at random (MCAR), missing at random (MAR) and missing not at random (NMAR). A variable is MCAR if the probability of its missing does not depend on the missing value of this variable or any other variable; a variable is MAR if the probability of its missing depends on other variables that have been fully recorded, but not on the values of unobserved variables; and a variable is NMAR if its probability of missing depends on information that has not been recorded, for example when a variable's missingness depends on its own value.(Little and Rubin, 2002)

Assume that we collect data on n subjects. For subject i, $i = 1, \ldots, n$, we observe a response variable Y_i, a vector $l_i = (l_{i1}, \ldots, l_{i(p+1-k)})'$ of $p + 1 - k$ covariates that are always fully observed, and a vector $m_i = (m_{i1}, \ldots, m_{ik})'$ of k covariates that may contain some missing components. We write $x_i = (l_i', m_i')'$, the vector of all $p + 1$ covariates. For each observation, we use an indicator variable R_i to denote if m_i is fully observed, that is, $R_i = 1$ if m_i is fully observed, and $R_i = 0$ otherwise. Let $t_i = (Y_i', l_i')' \in \mathbb{R}^t$, which is a vector of variables that are always observed. Then the aforementioned three types of missingness can be described as

$$\begin{aligned}
\text{(MCAR)} \quad & P(R_i = 1 \mid Y_i, x_i) && = P(R_i = 1), \\
\text{(MAR)} \quad & P(R_i = 1 \mid Y_i, x_i) && = P(R_i = 1 \mid t_i), \\
\text{(NMAR)} \quad & P(R_i = 1 \mid Y_i, x_i) && \text{No simplification.}
\end{aligned}$$

If the missing values are MCAR, then there is a loss in efficiency by dropping records with missing data, but this does not cause any systematic bias. This is because the probability of a subject having missing data is uniform across all subjects. Using the naive approach for mean or quantile regression will provide consistent estimators if the missingness is MCAR. The NMAR setting is much more challenging because the probability a values is missing depends on the unobserved variables. There are currently no techniques for NMAR that result in consistent estimates. Since MCAR only results in a loss of efficiency and finding unbiased estimates for NMAR data is not a tractable problem, in the missing data literature it is common to assume that the missing data are MAR. For the case of MAR we use the following notation

$$P(R_i = 1 \mid Y_i, x_i) = P(R_i = 1 \mid t_i) \equiv \pi_0(t_i) \equiv \pi_{i0}.$$

Consider the linear model (2.2). To handle the missing covariates when quantile regression is applied, a naive approach is to fit the model using only observations with complete data. The naive estimator is

$$\widehat{\beta}^N = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} R_i \rho_\tau(Y_i - x_i'\beta), \tag{4.1}$$

which is the standard quantile regression estimator only using subjects with complete data. For linear mean regression with covariates missing at random, it is known that the naive approach often leads to a biased estimator. This is also the case when we apply quantile regression naively to the observations with complete data only. To see this, we first observe that (4.1) implies that the estimator $\widehat{\beta}^N$ approximately solves the following estimating equation

$$G_n(\beta) = \sum_{i=1}^{n} R_i x_i \psi_\tau(Y_i - x_i'\beta) = 0, \tag{4.2}$$

where $\Psi_\tau(t) = \tau - I(t < 0)$ is the gradient function of $\rho_\tau(t)$. From a straightforward calculation, under the covariates missing at random assumption,

$$E\left[\sum_{i=1}^{n} R_i x_i \psi_\tau(Y_i - x_i'\beta)\right] = E\left[\sum_{i=1}^{n} \pi_{i0} x_i \psi_\tau(Y_i - x_i'\beta)\right].$$

Note that $E[\psi_\tau(Y_i - x_i'\beta)|x_i] = 0$. However since π_{i0} is a function of Y_i, it is not necessarily conditionally independent of $\psi_\tau(Y_i - x_i'\beta)$ given x_i. In general, we may not have $E\Big[\pi_{i0} x_i \psi_\tau(Y_i - x_i'\beta)\Big] = 0$, which is a necessary condition for $\widehat{\beta}^N$ to be consistent.

To alleviate the bias caused by missing data, we propose using the IPW approach. Let π_i be the probability that the ith data point is observed. The IPW method works by weighting the ith data point by R_i/π_i, note that records with missing data are assigned a weight of zero. IPW differs from the naive method by providing different weights to records with fully observed data. The intuition behind weighting is that for every fully observed data point with probability π_i of being fully observed, we expect $1/\pi_i$ data points with the same covariates if there was no missing data. For example a participant with complete data and $\pi_i = .25$ is given a weight of four. This is to account for the observed participant and the three participants with similar covariates who are likely to have incomplete data.(Tsiatis, 2006)

The weighted estimator approximately solves the weighted estimating equation

$$G_n^W(\beta) = \sum_{i=1}^{n} \frac{R_i}{\pi_{i0}} x_i \psi_\tau(Y_i - x_i'\beta) = 0. \tag{4.3}$$

For the intuition of why the weighted estimating equation is unbiased observe that

$$\begin{aligned} E\left[\frac{R_i}{\pi_{i0}} x_i \psi_\tau(Y_i - x_i'\beta)\right] &= E\left[\frac{\pi_{i0}}{\pi_{i0}} x_i \psi_\tau(Y_i - x_i'\beta)\right] \\ &= E\left[x_i E[\psi_\tau(Y_i - x_i'\beta)|x_i]\right] = 0. \end{aligned}$$

In practice, the missing data mechanism is often unknown and needs to be estimated, thus π_{i0} is replaced by $\hat{\pi}_i$. The weighted quantile regression estimator is formally defined as

$$\widehat{\beta}^W = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} \frac{R_i}{\hat{\pi}_i} \rho_\tau(Y_i - x_i'\beta). \tag{4.4}$$

As the above objective function is a weighted quantile objective function, the weighted quantile estimator can be easily computed using existing software. We address estimation and model selection for linear and additive partial linear models using IPW. To analyze the asymptotic behavior of the weighted estimators we need to impose conditions on π_{i0} and $\hat{\pi}_i$.

Condition 6

(Condition on the missing probability) There exists $\alpha_l > 0$ and $\alpha_u < 1$ such that $\alpha_l < \pi_{i0} < \alpha_u \, \forall i$. □

Condition 7

(Condition on the weights estimator) Assume a parametric form for π_{i0} with $\pi_{i0} \equiv \pi_i(\eta_0)$, $\hat{\pi}_i \equiv \pi_i(\hat{\eta})$ and $\hat{\eta}$ is the MLE of:

$$\prod_{i=1}^{n} \pi_i(\eta)^{R_i} (1 - \pi_i(\eta))^{(1-R_i)}$$

With conditions of asymptotic normality of $\hat{\eta}$ holding and $\left|\left|\frac{\partial \pi_i(\eta)}{\partial \eta}\right|\right|$ and $\left|\left|\frac{\partial^2 \pi_i(\eta)}{\partial \eta \partial \eta'}\right|\right|$ are bounded in a neighborhood of η_0. □

Condition 8

(Condition on asymptotic variances) For a matrix M, let $M > 0$ denote that M is a positive definite matrix.

- $\frac{1}{n}\sum_{i=1}^{n} f_i(0)x_i x_i' \xrightarrow{p} \tilde{\Sigma}_1 > 0$
- $\frac{1}{n}\sum_{i=1}^{n} \frac{\psi_\tau(\epsilon_i)^2}{\pi_i(\eta_0)} x_i x_i' \xrightarrow{p} \tilde{\Sigma}_2 > 0$
- $\frac{1}{n}\sum_{i=1}^{n} x_i \frac{1}{\pi_i(\eta_0)} \left(\frac{\partial \pi_i(\eta)}{\partial \eta}\right)'_{\eta=\eta_0} \psi_\tau(\epsilon_i) \xrightarrow{p} \tilde{\Sigma}_3$
- $\frac{1}{n}\sum_{i=1}^{n} \frac{\partial \pi_i(\eta_0)}{\partial \eta} \left(\frac{\partial \pi_i(\eta)}{\partial \eta}\right)'_{\eta=\eta_0} \frac{1}{\pi_i(\eta_0)(1-\pi_i(\eta_0))} \xrightarrow{p} I(\eta_0) > 0$ □

The lower bound from condition 6 is required to ensure that the weights do not increase to infinity while an upper bound is required because the asymptotic covariance of $\hat{\beta}^W$ depends on $\pi_{i0}(1-\pi_{i0})$. While condition 7 allows us to understand the asymptotic behavior of $\hat{\eta}$ which helps us analyze the asymptotics of $\hat{\beta}$. Condition 8 is needed to define the asymptotic variance of $\hat{\beta}$.

4.2 Linear Models

4.2.1 Estimation

First we consider the linear model with missing covariates and the weighted estimator of $\hat{\beta}^W$ from (4.4). Under some regulatory conditions the weighted estimator is asymptotically normal and unbiased.

Theorem 4.1

Let $\tilde{\Sigma}_m = \tilde{\Sigma}_2 - \tilde{\Sigma}_3 I(\eta_0)\tilde{\Sigma}_3'$ If conditions 1-2 and 6-8 hold then for $\hat{\beta}^W$ from (4.4)

$$\sqrt{n}(\hat{\beta}^W - \beta_0) \xrightarrow{d} N(0, \tilde{\Sigma}_1^{-1}\tilde{\Sigma}_m\tilde{\Sigma}_1^{-1}).$$ □

If the values of π_{i0} are known instead of estimated then the result from Theorem 4.1 changes to

$$\sqrt{n}(\hat{\beta}^W - \beta_0) \xrightarrow{d} N(0, \tilde{\Sigma}_1^{-1}\tilde{\Sigma}_2\tilde{\Sigma}_1^{-1}). \tag{4.5}$$

For symmetric matrices A and B let the notation $A \leq B$ mean $t'At \leq t'Bt$ for any vector $t \neq 0$ of appropriate dimension. Notice that

$$\tilde{\Sigma}_1^{-1}\tilde{\Sigma}_m\tilde{\Sigma}_1^{-1} \leq \tilde{\Sigma}_1^{-1}\tilde{\Sigma}_2\tilde{\Sigma}_1^{-1}.$$

Hence, it is asymptotically more efficient to use the estimated weights. The heuristic explanation is that the bias of the estimator comes from what is observed in the sample, not the population missingness generating mechanism. Therefore estimating the weights provides a more efficient estimator.(Robins et al., 1994).

4.2.2 Model Selection

Some covariates measured may not have a relationship with the response or fail to provide new information when conditioning on other variables. Determining which variables to include in the final model is a critical stage of analysis. Schwarz's BIC is a widely applied variable selection procedure. In the linear mean regression setting without missing data, it is known that under mild conditions choosing the model that minimizes BIC is model selection consistent. Meaning that if the true model is one of the candidates being considered then the true model with probability approaching one will have the smallest BIC value. When there is no missing data, BIC has been extended to quantile regression (Machado, 1993) and rank regression (Wang, 2009).

We write $x_i = (1, x_{i1}, \ldots, x_{i(p+1)})'$. We begin by indexing each candidate model by a $(p+1)$-dimensional binary vector $\nu = (1, \nu_1, \ldots, \nu_p)'$, where ν_j is one if the jth component of x_i belongs to the candidate model and is zero otherwise. The total number of ones in ν is denoted by d_ν, which describes the model complexity. Let $x_{i\nu}$ be the d_ν-dimensional subvector of x_i that contains the covariates in model ν; and let β_ν be the corresponding d_ν-dimensional subvector of parameters.

In the setting of quantile regression with missing covariates, the modified BIC for

the candidate model ν is defined as

$$\text{QBIC}_n(\nu) = \min_{\beta_\nu} \left\{ \sum_{i=1}^{n} \frac{R_i}{\pi_i(\hat{\eta})} \rho_\tau(Y_i - x'_{i\nu}\beta_\nu) + \frac{d_\nu \log n}{2} \right\}. \tag{4.6}$$

For model selection a new condition is required.

Condition 9

(Condition on misspecified models) If β_I is the limiting value for the estimator for an incorrect model, then for some positive constant

$$||\beta_I - \beta_0|| > C. \qquad \square$$

Condition 9 guarantees that asymptotically the objective function is minimized by the true model. This condition is used in the next theorem stating that minimizing (4.6) is a model selection consistent method.

Theorem 4.2

Assume that this class contains the true model, which is indexed by ν_0. Let the model selected by the modified BIC given in (4.6) be indexed by $\widehat{\nu}$, and assume that Conditions 1-2 and 6-9 are satisfied. Then as $n \to \infty$,

$$P(\widehat{\nu} = \nu_0) \to 1 \qquad \square$$

Therefore, the modified BIC for quantile regression with covariates missing at random possesses the property of model selection consistency. Sherwood et al. (2013) proposed a similar weighted version of BIC, but required the model of the weights could be modeled using logistic regression. The new theorem allows for model selection consistency to hold for other parametric formulations of π_i.

4.3 Additive Partial Linear Models

4.3.1 Estimation

Say there is an *i.i.d.* sample $\{Y_i, x_i, z_i\}_{i=1}^n$ with $Y_i \in \mathbb{R}$, $x_i \in \mathbb{R}^{p+1}$, including a constant, and $z_i \in \mathbb{R}^d$. We consider the additive partial linear quantile regression model

$$Y_i = x_i'\beta_0 + g_0(z_i) + \epsilon_i,$$

where $g_0(z_i) = \sum_{j=1}^d g_j(z_{ij})$ and $P(\epsilon_i < 0 \mid x_i, z_i) = \tau$, for some $\tau \in (0,1)$. A similar missing data mechanism is used, with $x_i = (l_i', m_i')'$ where l_i is a vector of covariates that is always observed and m_i is a vector of covariates that may contain missing values. Also $t_i = (Y_i, m_i', z_i')$ which is a vector of variables that are always observed. R_i remains the indicator variable for whether m_i contains complete data or not. We continue to use the MAR assumption, that is

$$P(R_i \mid x_i, Y_i, z_i) = P(R_i \mid t_i) = \pi_{i0}.$$

Basis splines are used to estimate the non-linear terms and $w(z_i)$ is the same basis vector defined in Chapter 3. The assumption that z_i is always observed avoids estimating the basis splines in the presence of missing data. The weighting method can also be used for the additive partial linear setting and we consider the estimates

$$\left(\widehat{\beta}_{PL}^W, \widehat{\gamma}^W\right) = \underset{(\beta,\gamma)}{\operatorname{argmin}} \sum_{i=1}^n \frac{R_i}{\pi_i(\hat{\eta})} \rho_\tau(Y_i - x_i'\beta - w(z_i)'\gamma). \tag{4.7}$$

Both $\hat{\beta}_{PL}^W$ and $\hat{\gamma}^W$ are consistent estimators and $\hat{\beta}_{PL}^W$ is asymptotically normal. For asymptotics we continue to use the relationship between X and Z stated in Section 3.2.

To establish results about $\left(\widehat{\beta}_{PL}^{W}, \widehat{\gamma}^{W}\right)$ we need a new condition similar to conditions 5 and 8.

Condition 10

(Condition on asymptotic variance)

- $\frac{1}{n}\sum_{i=1}^{n}\frac{\psi_\tau(\epsilon_i)^2}{\pi_i(\eta_0)}\delta_i\delta_i' \xrightarrow{p} \Sigma_2^W$,
- $\frac{1}{n}\sum_{i=1}^{n}\delta_i\frac{1}{\pi_i(\eta_0)}\left(\frac{\partial\pi_i(\eta)}{\partial\eta}\right)'_{\eta=\eta_0}\psi_\tau(\epsilon_i) \xrightarrow{p} \Sigma_3$. □

Theorem 4.3

If conditions 1-8 and 10 hold then for the estimators from (4.7)

$$\begin{aligned} ||\hat{\beta}_{PL}^{W} - \beta_0|| &= o_p(1), \\ \frac{1}{n}\sum_{i=1}^{n}(\hat{g}(z_i) - g_0(z_i))^2 &= O_p\left(\frac{dJ_n}{n}\right). \end{aligned}$$

Theorem 4.4

Let $\Sigma_m = \Sigma_2^W - \Sigma_3 I(\eta_0)\Sigma_3'$. If conditions 1-8 and 10 hold then for $\hat{\beta}_{PL}^{W}$ from (4.7)

$$\sqrt{n}(\hat{\beta}_{PL}^{W} - \beta_0) \xrightarrow{d} N(0, \Sigma_1^{-1}\Sigma_m\Sigma_1^{-1}).$$ □

Liang et al. (2004) considered a similar model for least squares estimation, but used a local linear kernel method to estimate the non-linear terms and did not allow for a subset of X to always have complete data. They also considered the augmented inverse probability weighting (AIPW) for which there currently is not an analog for quantile regression. In their work they found a similar asymptotic distribution for the IPW method.

4.3.2 Model Selection

Let $x_i = (x_i^{q\prime}, x_i^{c\prime})$ where $x_i^q \in \mathbb{R}^{q+1}$ and $x_i^c \in \mathbb{R}^{p-q}$ with the partial additive model we have used before of

$$\begin{aligned} Y_i &= x_i'\beta_0 + g_0(z_i) + \epsilon_i \\ &= x_i^{q\prime}\beta_{\mathbf{10}} + x_i^{c\prime}\beta_{\mathbf{20}} + g_0(z_i) + \epsilon_i, \end{aligned}$$

where $P(\epsilon_i \mid x_i, z_i) = \tau$. The difference now is that we assume some of the linear covariates do not have a relationship with the response. That is $\beta_{\mathbf{20}} = 0_{p-q}$ where 0_{p-q} is a $(p-q)$-dimensional vector of zeros. For model selection and estimation we minimize the following objective function,

$$(\hat{\beta}_{PL}^W(\lambda), \hat{\gamma}(\lambda)^W) = \underset{\beta,\gamma}{\operatorname{argmin}} \sum_{i=1}^{n} \frac{R_i}{\pi_i(\hat{\eta})} \rho_\tau(Y_i - x_i'\beta - w(z_i)'\gamma) + \sum_{j=1}^{p} p_\lambda(|\beta_j|). \quad (4.8)$$

Where $p_\lambda(\cdot)$ is a penalty function with tuning parameter λ. Penalized objective functions are a popular alternative to best subset model selection methods such as BIC. Penalized methods can be more computationally efficient than best subset methods, particularly when considering a large number of covariates. The L_1 penalty (LASSO), $p_\lambda(|\beta|) = \lambda|\beta|$ is a popular choice for penalized estimation.(Tibshirani, 1996) The L_1 penalty is known to over-penalize large coefficients and tends to be biased and requires strong conditions on the design matrix to achieve model selection consistency. This is usually not of concern if the goal is prediction, but can be undesirable if the goal is to identify the underlying model. Fan and Li (2001) proposed the SCAD penalty, motivated by finding a penalty function that provides an estimator with the oracle property, an estimator asymptotically equivalent to the estimator knowing which variables should be included in the model. For the SCAD penalty we

use the following function

$$p_\lambda(|\beta|) = \lambda|\beta| I(0 \leq |\beta| < \lambda) + \frac{a\lambda|\beta| - (\beta^2 + \lambda^2)/2}{a-1} I(\lambda \leq |\beta| \leq a\lambda) + \frac{(a+1)\lambda^2}{2} I(|\beta| > a\lambda), \text{ for some } a > 2.$$

Fan and Li (2001) recommended setting $a = 3.7$ and focusing on selecting λ. Figure 4.1 plots both the LASSO and SCAD penalty functions for $\lambda = 2$ and $a = 3.7$ for the SCAD penalty function. One appeal of the SCAD penalty is that it does not over-penalize large coefficients. A consequence of this property is the penalty function is not convex and therefore minimizing (4.8) is not a convex minimization problem and a local minimum is not guaranteed to be a global minimum. Current theory and estimation methods are limited to finding local minimums of (4.8) when $p_\lambda(\cdot)$ is non-convex. For both penalty functions, the tuning parameter λ controls the complexity of the selected model and goes to zero as n increases to ∞. A more thorough presentation of the SCAD penalty is provided in Chapter 5 where we consider using the SCAD penalty when $p >> n$.

Fan and Li (2001) proposed using the SCAD penalty in the least squares setting and suggested it could be used for robust methods, such as median regression. Wu and Liu (2009) studied using the SCAD penalty for variable selection of quantile regression models. Liu et al. (2011) proposed using the SCAD penalty for variable selection of the linear components of a partial linear model. To the best of our knowledge nobody has investigated the use of the SCAD penalty with an additive partial linear quantile regression model. Another novel contribution is using the SCAD penalty with the weighted objective function for variable selection in the presence of missing data.

For the case with missing covariates we formally define The oracle estimator for

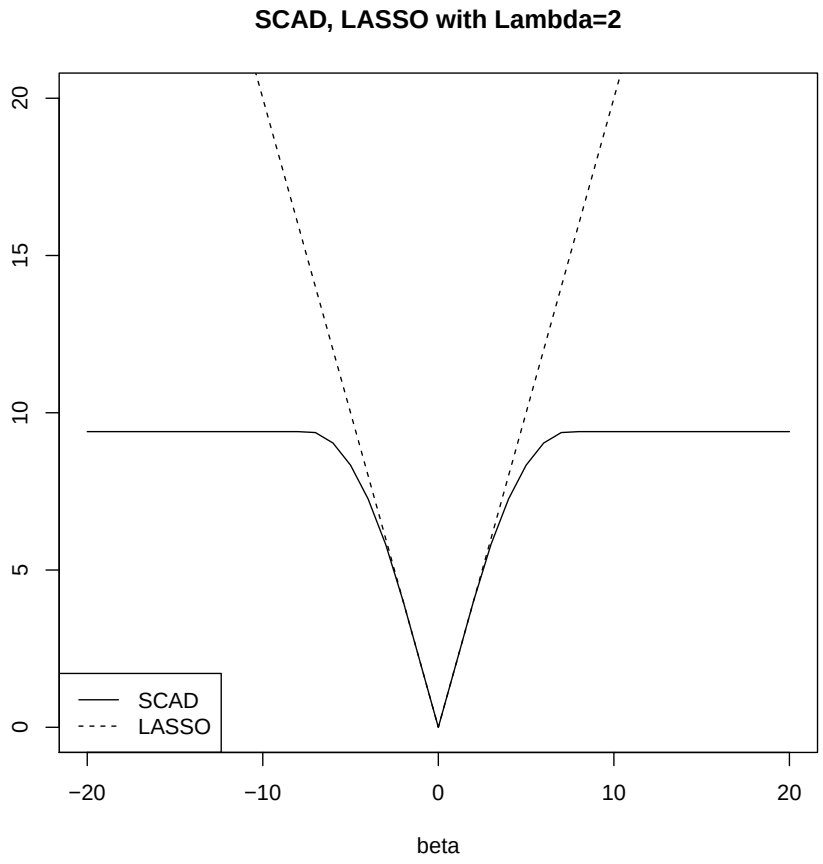

Figure 4.1: SCAD and LASSO plots

$(\beta_0', \gamma_0')'$ as $\left(\tilde{\beta}_{PL'}^{W}, \tilde{\gamma}^{W'}\right)'$, where $\tilde{\beta} \equiv \left(\hat{\beta}_{\mathbf{PL_1}}^{\mathbf{W}'}, \mathbf{0}_{p-q}'\right)'$ and

$$\left(\hat{\beta}_{\mathbf{PL_1}}^{\mathbf{W}}, \tilde{\gamma}^{W}\right) = \underset{(\beta_{\mathbf{1}}, \gamma)}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \frac{R_i}{\pi_i(\hat{\eta})} \rho_\tau(Y_i - x_i' \left(\beta_{\mathbf{1}}', \mathbf{0}_{p-q}'\right)' - w(z_i)'\gamma). \tag{4.9}$$

The oracle estimator sets to zero the coefficients for any linear covariates that do not have a relationship with the response. Otherwise the estimator is similar to the estimator from (4.7) and its asymptotic properties follow from Theorem 4.3 and Theorem 4.4. Our next theorem states that asymptotically the oracle estimator is equivalent to a local minimum estimator of (4.8) using the SCAD penalty function.

Theorem 4.5

Assume conditions 1-8 and 10 are satisfied. Let $\mathcal{E}_n(\lambda)$ be the set of local minima of the the penalized objective function from (4.8) using the SCAD penalty function with tuning parameter λ. Let $\hat{\eta} \equiv \left(\hat{\beta}_{\mathbf{PL_1}}^{\mathbf{W}}, \tilde{\gamma}^{W}\right)$ be the oracle estimator from (4.9). If $n^{-1/2} J_n = o(\lambda)$ then as $n \to \infty$

$$P\left(\hat{\eta} \in \mathcal{E}_n(\lambda)\right) \to 1. \qquad \square$$

4.4 Simulations

Monte Carlo simulations were performed to evaluate the finite sample size performance of the estimators. In the first simulation setting we focus on estimation. Comparing the weighted method to the naive method and a full data method, which would be the estimator if the values of the missing data were known. In application this is not possible, but it provides a comparison point for the weighted estimator.

In the second simulation we take into account variable selection for the parametric component of the additive partial linear model. Again, we compare the weighted and

naive methods. We also compare performance for the SCAD and LASSO penalty. For the SCAD penalty we are interested in verifying that local minimums of the penalized objective function are good estimators.

4.4.1 Estimation

Define $g_1(z) = \sin(2\pi z)$ and $g_2(z) = x^3 - .25$. The model is

$$Y_i = -3 + x_{i1} - x_{i2} + x_{i3} + g_1(z_{i1}) + g_2(z_{i2}) + \epsilon_i$$

where $P(\epsilon_i \leq 0 \mid x_i, z_i) = \tau$, $x_1 \sim N(0,1)$, $x_2 \sim N(0,1)$, $x_3 \sim \text{Ber}(.5)$, $Z_1 \sim U[0,1]$ and $Z_2 \sim U[0,1]$. We consider three different settings for ϵ_i: (1) $\epsilon_i \sim N(0,1)$; (2) $\epsilon_i \sim T_3$; and (3) $\epsilon_i \sim (1 + x_{i3})z_i$ where $z_i \sim N(0,1)$. For the first and second case we fit a model for $\tau = .5$. For the heteroscedastic case we fit models for $\tau = .7$ because modeling non-central quantiles provides insight in this case that is lost by only focusing on central behavior.

The missing model is

$$\text{logit}(P(R_i) = 1) = 4 + Y_i + x_{i2} + z_{i1} - z_{i2}.$$

Weights in the model are estimated by first fitting a logistic regression with R_i as the response and Y_i, x_{i2}, z_{i1} and z_{i2} as predictors. The weights are the inverse of the fitted values from this model. For estimation of β_0 and g_0 three models are considered: (1) Weighted: estimates using the IPW method; (2) Naive: records with missing values are dropped from the analysis, no weighting done to account for missing values; (3) Full: standard quantile regression analysis using known values for the missing data.

Two hundred and fifty simulations were performed for each setting. Let $\hat{\beta}^k$, $\hat{g}^k$ and r_n^k denote the linear estimate, non-linear estimate and number of fully observed

subjects of the kth simulation. Simulations are summarized using the following statistics

1. Bias: $\sum_{j=0}^{3} \frac{1}{250} \left| \sum_{k=1}^{250} \widehat{\beta}_j^k - \beta_{j0} \right|$,
2. MSE: $\frac{1}{250} \sum_{j=0}^{3} \sum_{k=1}^{250} \left[\widehat{\beta}_j^k - \beta_{j0} \right]^2$,
3. AADE: $\frac{1}{250} \sum_{k=1}^{250} \frac{1}{n} \sum_{i=1}^{n} \left| \hat{g}^k(z_i) - g_0(z_i) \right|$,
4. Average r_n: $\frac{1}{250} \sum_{k=1}^{250} r_n^k$.

A weight threshold of 50 was used, that is, any observations that had a weight of over 50 were assigned a weight of 50. This avoids the case of a very large weight being assigned to one observation which typically results in poor estimators. The threshold also relates to condition 6 which assumes that there is a lower bound to the probability that a subject would have complete data. To select the number of basis functions we use a similar approach to the simulations in Chapter 3 that also accounts for the weighted objective function. For the weighted method define $\hat{\beta}_{PL}^W(J_{1_n}, J_{2_n})$ and $\hat{\gamma}^W(J_{1_n}, J_{2_n})$ be the estimates derived when using J_{1_n} and J_{2_n} basis functions. Further let $\nu(J_{1_n}, J_{2_n}) = 4 + J_{1_n} + J_{2_n}$, the degrees of freedom in a model with J_{n_1} and J_{n_2}. Let

$$\begin{aligned} QBIC^W(J_{1_n}, J_{2_n}) &= \log\left(\sum_{i=1}^{n} \frac{R_i}{\pi_i(\hat{\eta})} \rho_\tau \left(Y_i - x_i' \hat{\beta}^W(J_{1_n}, J_{2_n}) - w(z_i)' \hat{\gamma}^W(J_{1_n}, J_{2_n}) \right) \right) \\ &+ \frac{\nu(J_{1_n}, J_{2_n}) \log(n)}{2n}. \end{aligned}$$

The final model is selected by finding the combination of J_{1_n} and J_{2_n} that minimizes $QBIC^W(\cdot)$. For the naive method the weights of $\frac{R_i}{\pi_i(\hat{\eta})}$ are replaced with R_i, while for the full data method uses $QBIC(\cdot)$ proposed in the simulations section of Chapter 3. Results of the simulations are presented in Table 4.1-Table 4.3.

Method	n	Average r_n	Bias	MSE	AADE
Naive	100	72	0.57	0.38	0.32
Naive	300	217	0.59	0.24	0.23
Naive	1000	723	0.59	0.19	0.19
Full	100	100	0.04	0.15	0.24
Full	300	300	0.03	0.05	0.15
Full	1000	1000	0.01	0.01	0.09
Weighted	100	72	0.02	0.52	0.69
Weighted	300	217	0.15	0.20	0.46
Weighted	1000	723	0.17	0.07	0.33

Table 4.1: Missing Additive Partial Linear Simulation Results for $\epsilon_i \sim N(0,1)$

Method	n	Average r_n	Bias	MSE	AADE
Naive	100	71	0.73	0.52	0.37
Naive	300	213	0.70	0.31	0.24
Naive	1000	711	0.72	0.28	0.20
Full	100	100	0.02	0.19	0.28
Full	300	300	0.03	0.05	0.16
Full	1000	1000	0.02	0.02	0.10
Weighted	100	71	0.08	0.94	0.90
Weighted	300	213	0.20	0.35	0.52
Weighted	1000	711	0.19	0.11	0.39

Table 4.2: Missing Additive Partial Linear Simulation Results for $\epsilon_i \sim T_3$

The bias of the naive method is clear in all three settings. An interesting result is that for larger sample sizes the bias of the weighted method stabilizes, but is actually larger than for $n = 100$. This is a consequence of using the thresholding method for the weights. This approach does cause some bias, but drastically reduces the variance of the weighted estimators. The larger the sample size the more likely the thresholding method needs to be used resulting in the weighted method having larger bias for larger sample sizes. The weighted methods have larger variances, but for larger sample sizes is noticeably less biased than the naive method. Thus for larger sample sizes the weighted method has a smaller MSE than the naive method. In all settings estimation of the non-linear functions improves as n increases. The naive

Method	n	Average r_n	Bias	MSE	AADE
Naive	100	70	0.68	0.55	0.45
Naive	300	210	0.71	0.32	0.29
Naive	1000	702	0.70	0.23	0.22
Full	100	100	0.06	0.30	0.35
Full	300	300	0.05	0.10	0.21
Full	1000	1000	0.01	0.03	0.12
Weighted	100	70	0.10	1.09	1.00
Weighted	300	210	0.17	0.37	0.64
Weighted	1000	702	0.22	0.12	0.41

Table 4.3: Missing Additive Partial Linear Simulation Results for ϵ_i heteroscedastic

method outperforms the weighted method for estimating the non-linear functions. In our next simulation setting the weighted method performs better.

4.4.2 Model Selection

For the model selection simulations we consider a similar model, but consider extra covariates which are not part of the true model. Generate $\tilde{X} \sim N_7(0, \Sigma)$ where $\Sigma_{ij} = .5^{|i-j|}$, $X_8 \sim \text{Ber}(.5)$, $Z_1 \sim U[0,1]$ and $Z_2 \sim U[-1,1]$. Further let $X = [\tilde{X}\, X_8] \in \mathbb{R}^{n\times 8}$. Define $g_1(z) = \sin(2\pi z)$ and $g_2(z) = x^3$. The data generating mechanism is

$$Y_i = x_{i1} - x_{i3} + x_{i8} + g_1(z_{i1}) + g_2(z_{i2}) + \epsilon_i.$$

We considered three different settings for ϵ_i similar to those that were used in the estimation simulations: (1) $\epsilon_i \sim N(0,1)$; (2) $\epsilon_i \sim T_3$; and (3) $\epsilon_i \sim (1 + x_{i8})\xi_i$ where $\xi_i \sim N(0,1)$. For the heteroscedastic case we modeled for $\tau = .7$, while modeling the conditional median for the other two settings.

All X variables may have missing data except for X_3. The missing model is

$$\text{logit}(P(R_i) = 1) = 1 + Y_i + x_{i3} - z_{i1} + z_{i2}.$$

Four methods of estimation are considered: (1) "SCAD Full" which uses the SCAD penalized objective function with no missing data; (2) "SCAD Naive" which uses the SCAD penalized objective function and drops all records with missing data; (3) "SCAD Wt" which uses the SCAD penalty with the IPW objective function; (4) "LASSO Wt" which uses the LASSO penalty with the IPW objective function. In all simulations a weight threshold of 50 was used to prevent any single observation from having excessive weight in the analysis. Along with reporting r_n, $Bias$, MSE and $AADE$ as defined in the previous section we report an additional three summary statistics for model selection:

1. TV: average number of linear covariates correctly included in the model,

2. False Variables (FV): average number of linear covariates incorrectly included in the model,

3. True: proportion of times the true model is exactly identified.

In these simulations we considered the number of basis functions to use and the choice of λ. For both nonlinear variables we consider 3 to 15 basis functions. Let $\nu = \nu_1 + J_{1n} + J_{2n}$ where J_{n1} and J_{n2} were defined in the estimation simulations and ν_1 is the number of parametric terms included in the model. Let $\hat{\beta}_\lambda(J_{1n}, J_{2n})$ and $\hat{\gamma}_\lambda(J_{1n}, J_{2n})$ be the fits for a given λ, J_{1n} and J_{2n}. For the SCAD Wt method we choose λ, J_{1n} and J_{2n} which minimizes

$$\begin{aligned} QBIC^W(\lambda, J_{1n}, J_{2n}) &= \log\left(\sum_{i=1}^{n} \frac{R_i}{\pi_i(\hat{\eta})}\rho_\tau(Y_i - x_i'\hat{\beta}_\lambda(J_{1n}, J_{2n}) - w(z_i)'\hat{\gamma}_\lambda(J_{1n}, J_{2n}))\right) \\ &+ \frac{\nu \log(n)}{2n}. \end{aligned}$$

For "SCAD Naive" we replace the weights of $\frac{R_i}{\pi_i(\hat{\eta})}$ with R_i, while for "SCAD Full" a

full data version is used without any weights. For all of the SCAD based methods we set $a = 3.7$ as suggested in Fan and Li (2001). The LASSO penalty is more appropriate when prediction is the problem of interest and for this reason we use 5-folds cross-validation to select λ, J_{n1} and J_{n2} for the "LASSO Wt" method.

In the simulations section of Chapter 5 we present an algorithm for how to solve the penalized objective function for both SCAD and LASSO. Simulations were run with sample sizes of 200, 400 and 1000. For each setting 160 simulations were performed and results are reported in Table 4.4-Table 4.6.

All methods improved with sample size as expected from our asymptotic results. The four approaches worked well in selecting the three covariates that are part of the true model. The "LASSO Wt" method tends to select a larger model, which is expected because this is typical behavior for LASSO methods and cross validation was used to select the tuning parameter. The SCAD based methods tend to produce a smaller model and have a higher probability of selecting the true model as we would expect from the theory. A surprising result is that the "SCAD Wt" method tends to select a larger model than "SCAD Naive". We believe this is caused by the extra variability introduced by the weighted method. The extra variability from the weighted method decreases with sample size, but the bias from the naive method remains. The advantage of the weighted method can be seen when comparing Bias, MSE and AADE of the naive and weighted methods.

Method	n	r_n	TV(3)	FV	True	Bias	MSE	AADE
SCAD Full	200	200	3.00	0.01	0.99	0.09	0.06	0.26
SCAD Full	400	400	3.00	0.00	1.00	0.06	0.03	0.20
SCAD Full	1000	1000	3.00	0.00	1.00	0.03	0.01	0.15
SCAD Naive	200	129	2.96	0.02	0.95	0.31	0.15	0.41
SCAD Naive	400	259	3.00	0.00	1.00	0.27	0.07	0.35
SCAD Naive	1000	647	3.00	0.00	1.00	0.24	0.04	0.33
SCAD Wt	200	129	2.96	0.44	0.66	0.17	0.20	0.39
SCAD Wt	400	259	2.99	0.12	0.88	0.06	0.09	0.30
SCAD Wt	1000	647	3.00	0.01	0.99	0.04	0.04	0.21
LASSO Wt	200	129	2.98	2.61	0.01	0.73	0.39	0.39
LASSO Wt	400	259	3.00	2.18	0.06	0.59	0.25	0.30
LASSO Wt	1000	647	3.00	1.36	0.20	0.57	0.17	0.25

Table 4.4: Missing Data Additive Partial Linear Simulation Results for $\tau = 0.5$ and $\epsilon \sim N(0, 1)$

Method	n	r_n	TV(3)	FV	True	Bias	MSE	AADE
SCAD Full	200	200	2.99	0.00	0.99	0.08	0.07	0.27
SCAD Full	400	400	3.00	0.00	1.00	0.07	0.03	0.22
SCAD Full	1000	1000	3.00	0.00	1.00	0.05	0.01	0.16
SCAD Naive	200	128	2.96	0.02	0.94	0.37	0.17	0.49
SCAD Naive	400	256	2.99	0.00	0.99	0.34	0.10	0.47
SCAD Naive	1000	641	3.00	0.00	1.00	0.31	0.06	0.43
SCAD Wt	200	128	2.91	0.40	0.65	0.24	0.30	0.47
SCAD Wt	400	256	2.99	0.16	0.86	0.09	0.12	0.35
SCAD Wt	1000	641	3.00	0.05	0.96	0.06	0.05	0.26
LASSO Wt	200	128	2.92	2.71	0.01	0.85	0.54	0.45
LASSO Wt	400	256	2.99	2.17	0.06	0.66	0.32	0.36
LASSO Wt	1000	641	3.00	1.62	0.17	0.61	0.21	0.28

Table 4.5: Missing Data Additive Partial Linear Simulation Results for $\tau = 0.5$ and $\epsilon \sim T_3$

Method	n	r_n	TV(3)	FV	True	Bias	MSE	AADE
SCAD Full	200	200	3.00	0.03	0.97	0.09	0.13	0.58
SCAD Full	400	400	3.00	0.00	1.00	0.06	0.06	0.55
SCAD Full	1000	1000	3.00	0.00	1.00	0.05	0.02	0.55
SCAD Naive	200	126	3.00	0.12	0.89	0.32	0.23	0.84
SCAD Naive	400	251	3.00	0.01	0.99	0.34	0.13	0.81
SCAD Naive	1000	630	3.00	0.00	1.00	0.37	0.09	0.82
SCAD Wt	200	126	2.98	0.51	0.64	0.16	0.43	0.68
SCAD Wt	400	251	2.99	0.19	0.84	0.09	0.20	0.61
SCAD Wt	1000	630	3.00	0.01	0.99	0.02	0.05	0.56
LASSO Wt	200	126	2.98	2.39	0.03	0.98	0.74	0.79
LASSO Wt	400	251	3.00	1.85	0.12	0.87	0.52	0.75
LASSO Wt	1000	630	3.00	1.02	0.36	0.79	0.32	0.72

Table 4.6: Missing Data Additive Partial Linear Simulation Results for $\tau = 0.7$ and ϵ heteroscedastic

4.5 Applied Example: Medical Cost Data

The overall cost of health care is driven by high cost patients. To determine effective strategies for controlling health care costs we need to directly model these high cost patients. Sherwood et al. (2013) proposed using the weighted quantile regression objective function to model health care costs, but assumed a linear relationship between the log of health care cost and the observed predictors. In this analysis we use the more flexible additive partial linear model and use the penalized objective functions for model selection.

The data we analyze came from a clinical study on the cost-effectiveness of a computer-assisted prospective drug utilization review program presented in Tierney et al. (1995). The study was conducted in the primary care system of Indiana University Medical Group Primary Care. The data set was analyzed in Zhou et al. (2001) using a heteroscedastic mean regression model. In their analysis, patients with missing information have been excluded. This data set has the following variables:

1. Charge ($): Amount charged for the health care provided,
2. Age: Age of the patient,
3. African-American: Binary variable indicating if the patient is African-American (1) or not (0),
4. Female: Binary variable indicating if the patient is female (1) or not (0),
5. Education: Years of education,
6. Live Alone: Binary variable indicating if the patient lives alone (1) or not (0),
7. Doctor Satisfaction: Rating of the patients satisfaction of their doctor on a scale of 1-5,

8. Pharmacist Satisfaction: Rating of the patients satisfaction of their pharmacist on a scale of 1-5,

9. SF-36 Phys: Measurement of physical fitness on a scale of 0-100,

10. SF-36 GH: Measurement of general health on a scale of 0-100,

11. Bad Timing: Did the patient take medicine as scheduled (1) or not (0),

12. Bad Reaction: Binary variable indicating if the patient stopped taking medicine because of a bad reaction (1) or not (0),

13. Sexually Active: Binary variable indicating if the patient is sexually active (1) or not (0).

There are 712 patients in the data set and 95 patients with missing data, about 13% of the records. The variables that have missing values are Doctor Satisfaction, Pharmacist Satisfaction, Education, SF-36 Phys and SF-36 GH. In our analysis we use a log-transformed "charge" variable. There are 17 patients with zero charges. To accommodate the log transformation patients with a zero charge are assigned a charge of 5 dollars, smaller than the smallest non-zero charge of $12. Mean regression could be sensitive to these changes, but the estimates of these conditional quantiles are robust to small changes of the response in the lower tail. Unlike mean regression the transformation of the quantiles can easily be interpreted, that is the conditional median of the logged charges is the log of the conditional median of charges.

With a small percentage of patients accounting for most of the health care costs, it is of particular interest to consider the patients with high costs, in other words, the high conditional quantiles, such as $\tau = 0.8$ and 0.9. We also model the conditional median to understand central tenancies. To account for the missing data we fit a logistic regression using the missing data indicator as a response variable and all

of the fully observed variables, with charge still on the log scale, as predictors. A summary of this model is provided in Table 4.7 which shows that the important predictor for missingness is the cost of the patient. In this data set high cost patients are less likely to have missing data.

	Estimate	Std. Error	z value	Pr(>|z|)
Intercept	-0.3021	0.8255	-0.37	0.7144
log(Charge)	0.3488	0.0658	5.30	0.0000
Age	-0.0069	0.0109	-0.64	0.5237
African-American	0.1837	0.2336	0.79	0.4318
Female	0.0995	0.2581	0.39	0.6999
Live Alone	-0.2216	0.2548	-0.87	0.3844
Bad Timing	0.3697	0.3481	1.06	0.2882
Bad Reaction	-0.3231	0.3802	-0.85	0.3953
Sexually Active	-0.1231	0.2487	-0.49	0.6206

Table 4.7: Logistic Regression Model for Missingness in Cost Data

Next we fit models using the SCAD Weighted, SCAD Naive and LASSO Weighted methods outlined in the previous section. We model age as having a non-linear relationship with the response and consider all other variables as linear variables. Tuning parameter and number of basis coefficients used with the SCAD penalty were determined by minimizing $QBIC^W(\lambda, J_n)$ for the SCAD weighted method, an unweighted version is used for the SCAD Naive approach. Five-folds cross validation was used with the LASSO penalty. In addition to these models we also consider a naive and weighted saturated model, "Sat Naive" and "Sat Wt"' respectively. For these models age is fit as a non-linear predictor and all other variables are included as linear predictors. The weighted saturated model uses the weighted objective function to handle the missing data, while the naive saturated model drops all records with missing data and does not account for the missing data. Estimates for the models are presented in Table 4.8-Table 4.10. A value of * indicates that the coefficient was not included in the model.

The models change depending on the method and τ. The variable SF36_PF is included in all of the models for $\tau = .5$, the only case where a variable is present in all of the models for a given τ. The bad reaction indicator variable may be an important variable for high cost patients. It is included in both the naive and weighted SCAD models for $\tau = .8$ and the weighted method for $\tau = .9$. The data was originally collected to determine if pharmacists satisfaction can help lower health care costs. There is little evidence of that being the case, with "Pharm_Sat" only included in the median model using the LASSO objective function with weights.

Method	LASSO Wt	Sat Naive	SAT Wt	SCAD Wt	SCAD Naive
Intercept	7.60	6.71	6.72	6.21	7.39
AA	*	-0.11	-0.13	*	*
Female	*	-0.20	-0.13	*	*
Dr. Sat	*	-0.08	-0.06	*	*
Pharm Sat	-0.03	-0.16	-0.20	*	*
Alone	*	0.13	0.10	*	*
Education	*	0.05	0.04	*	*
SF36_PF	-0.22	-0.22	-0.22	-0.19	-0.06
SF36_GH	-0.16	-0.20	-0.24	*	*
Bad Timing	*	-0.19	-0.16	*	*
Bad React	*	0.37	0.35	*	*
Sexually Active	*	-0.24	-0.26	*	*

Table 4.8: Median Health Care Cost Models

We used a random partition method to calculate the predicative performance of the different models. Six hundred and twelve patients are randomly selected for training data and the remaining 100 patients are used as testing data. We fit all five methods for $\tau = 0.5$, 0.8 and 0.9. Let r_n be the number of records with complete data from the testing data. Then we apply the selected model and the full model to those data points with complete records in the testing data, and evaluate their predictive performance by calculating the mean absolute prediction error $r_n^{-1} \sum_{j=1}^{r_n} \rho_\tau(\widehat{Y}_j - Y_j)$, where $\widehat{Y}_j$ is the predicted value for the jth patient with complete data. We repeat the

Method	LASSO Wt	Sat Naive	SAT Wt	SCAD Wt	SCAD Naive
Intercept	10.49	10.55	9.30	12.28	7.90
AA	*	-0.32	-0.26	*	*
Female	*	-0.41	-0.32	*	*
Dr. Sat	*	-0.02	-0.05	*	*
Pharm Sat	*	0.04	0.04	*	*
Alone	*	0.47	0.45	*	*
Education	*	0.02	-0.02	*	*
SF36_PF	-0.18	-0.23	-0.24	*	*
SF36_GH	-0.05	-0.14	-0.19	*	*
Bad Timing	*	-0.48	-0.44	-0.45	*
Bad React	*	0.85	0.86	0.86	0.77
Sexually Active	*	-0.35	-0.30	*	-0.53

Table 4.9: .8 Quantile Health Care Cost Models

above random partition 500 times and report the overall mean absolute prediction error for each model. The results are summarized in Table 4.11. We observe that the selected sparse models have similar predictive performance comparing to the full model. Hence the SCAD penalty effectively reduces the model complexity without sacrificing the predictive ability. The difference between the weighted methods and naive methods is small. Suggesting that any bias due to missing data is small.

Method	LASSO Wt	Sat Naive	SAT Wt	SCAD Wt	SCAD Naive
Intercept	11.82	12.10	12.35	10.14	9.97
AA	*	-0.31	-0.35	-0.80	-0.72
Female	*	-0.69	-0.63	*	-0.49
Dr. Sat	*	-0.15	-0.14	*	*
Pharm Sat	*	-0.11	-0.15	*	*
Alone	*	0.56	0.60	0.55	*
Education	*	0.12	0.11	*	*
SF36_PF	-0.06	-0.28	-0.22	*	*
SF36_GH	*	-0.32	-0.34	*	*
Bad Timing	*	-0.48	-0.46	*	*
Bad React	*	0.84	0.76	0.60	*
Sexually Active	*	-0.06	-0.04	*	*

Table 4.10: .9 Quantile Health Care Cost Models

τ	LASSO WT	Sat Naive	Sat Wt	SCAD NAIVE	SCAD Wt
0.5	0.53	0.52	0.53	0.55	0.54
0.8	0.40	0.39	0.39	0.41	0.41
0.9	0.26	0.25	0.25	0.27	0.27

Table 4.11: Random Partition Results for Modeling Healthcare Cost

4.6 Proofs

Proof of Theorem 4.1

Proof: Define $\theta = \sqrt{n}(\beta - \beta_0)$ and $\hat{\theta} = \sqrt{n}(\hat{\beta}^W - \beta_0)$. Note that

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{n} \frac{R_i}{\pi_i(\hat{\eta})} \rho_\tau(\epsilon_i - n^{-1/2} x_i' \theta) - \rho_\tau(\epsilon_i).$$

Using Knight's Identity (Knight, 1998)

$$\begin{aligned}\sum_{i=1}^{n}\frac{R_i}{\pi_i(\hat{\eta})}\rho_\tau(\epsilon_i - n^{-1/2}x_i'\theta) - \rho_\tau(\epsilon_i) &= -\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{R_i}{\pi_i(\hat{\eta})}x_i'\theta\psi_\tau(\epsilon_i) \\ &+ \sum_{i=1}^{n}\frac{R_i}{\pi_i(\hat{\eta})}\int_0^{x_i'\theta n^{-1/2}}[I(\epsilon_i \le s) - I(\epsilon_i \le 0)]\,ds \\ &\equiv A_{n1} + A_{n2}.\end{aligned}$$

Where the definitions of A_{n1} and A_{n2} are the separate sums obtained by using Knights identity. First for A_{n1}:

$$\begin{aligned}A_{n1} &= -\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{R_i}{\pi_i(\eta_0)}x_i'\theta\psi_\tau(\epsilon_i) \\ &\quad -\frac{1}{\sqrt{n}}\sum_{i=1}^{n}R_i\left(\frac{1}{\pi_i(\hat{\eta})} - \frac{1}{\pi_i(\eta_0)}\right)x_i'\theta\psi_\tau(\epsilon_i) \\ &\equiv A_{n11} + A_{n12}.\end{aligned}$$

Using Taylor expansion, condition 7 and theory regarding asymptotic normality of MLEs

$$\begin{aligned}A_{n12} &= -\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{1}{\pi_i(\eta_0)^2}x_i'\theta\psi_\tau(\epsilon_i)\left(\frac{\partial\pi_i(\eta)}{\partial\eta}\right)'_{\eta=\eta_0}(\hat{\eta}-\eta_0) + o_p(1) \\ &= -\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\pi_i(\eta_0)^2}x_i'\theta\psi_\tau(\epsilon_i)\left(\frac{\partial\pi_i(\eta)}{\partial\eta}\right)'_{\eta=\eta_0}I(\eta_0)^{-1} \\ &\times\ -n^{-1/2}\sum_{i=1}^{n}\left(\frac{\partial\pi_i(\eta)}{\partial\eta}\right)_{\eta=\eta_0}\left(\frac{R_i-\pi_i(\eta_0)}{\pi_i(\eta_0)(1-\pi_i(\eta_0)}\right) + o_p(1) \\ &= \theta'\tilde{\Sigma}_3 I(\eta_0)^{-1}n^{-1/2}\sum_{i=1}^{n}\left(\frac{\partial\pi_i(\eta)}{\partial\eta}\right)_{\eta=\eta_0}\left(\frac{R_i-\pi_i(\eta_0)}{\pi_i(\eta_0)(1-\pi_i(\eta_0))}\right) + o_p(1)\end{aligned}$$

Using the asymptotically equivalent version of A_{n12}

$$\begin{aligned} A_{n1} &= n^{-1/2}\sum_{i=1}^{n}\left[\frac{R_i}{\pi_i(\eta_0)}x_i'\theta\psi_\tau(\epsilon_i) - \theta'\tilde{\Sigma}_3 I(\eta_0)^{-1}\left(\frac{\partial\pi_i(\eta)}{\partial\eta}\right)_{\eta=\eta_0}\left(\frac{R_i-\pi_i(\eta_0)}{\pi_i(\eta_0)(1-\pi_i(\eta_0))}\right)\right] \\ &\quad +o_p(1). \end{aligned}$$

The sum is of mean zero random variables. We check the variance and covariance of the two sums.

$$\mathrm{Var}\left(\frac{R_i}{\pi_i(\eta_0)}x_i'\theta\psi_\tau(\epsilon_i)\right) = \theta' E\left[\frac{1}{\pi_i(\eta_0)}x_i x_i'\psi_\tau(\epsilon_i)^2\right]\theta = \theta'\tilde{\Sigma}_2\theta.$$

For the variance of the second sum:

$$\begin{aligned} &\mathrm{Var}\left(\theta'\tilde{\Sigma}_3 I(\eta_0)^{-1}\left(\frac{\partial\pi_i(\eta)}{\partial\eta}\right)_{\eta=\eta_0}\frac{R_i-\pi_i(\eta_0)}{\pi_i(\eta_0)(1-\pi_i(\eta_0))}\right) \\ =\ &\theta'\tilde{\Sigma}_3 I(\eta_0)^{-1}E\left[\left(\frac{\partial\pi_i(\eta)}{\partial\eta}\right)_{\eta=\eta_0}\left(\frac{\partial\pi_i(\eta)}{\partial\eta}\right)'_{\eta=\eta_0}\left(\frac{R_i-\pi_i(\eta_0)}{\pi_i(\eta_0)(1-\pi_i(\eta_0))}\right)^2\right] I(\eta_0)^{-1}\tilde{\Sigma}_3'\theta \\ =\ &\theta'\tilde{\Sigma}_3 I(\eta_0)^{-1}\tilde{\Sigma}_3'\theta. \end{aligned}$$

For the covariance:

$$\begin{aligned} &E\left[\frac{R_i}{\pi_i(\eta_0)}x_i'\theta\psi_\tau(\epsilon_i)\left(\frac{\partial\pi_i(\eta)}{\partial\eta}\right)'_{\eta=\eta_0}\frac{R_i-\pi_i(\eta_0)}{\pi_i(\eta_0)(1-\pi_i(\eta_0))}I(\eta_0)^{-1}\tilde{\Sigma}_3'\theta\right] \\ =\ &\theta' E\left[x_i\psi_\tau(\epsilon_i)\left(\frac{\partial\pi_i(\eta)}{\partial\eta}\right)'_{\eta=\eta_0}\frac{R_i}{\pi_i(\eta_0)}\right] I(\eta_0)^{-1}\tilde{\Sigma}_3'\theta \\ =\ &\theta'\tilde{\Sigma}_3 I(\eta_0)^{-1}\tilde{\Sigma}_3'\theta. \end{aligned}$$

Then by CLT, for $Z \sim N(0, \tilde{\Sigma}_2 - \tilde{\Sigma}_3 I(\eta_0)^{-1}\tilde{\Sigma}_3')$

$$A_{n1} \xrightarrow{d} \theta' Z\theta.$$

We use a similar separation for A_{n2}.

$$\begin{aligned} A_{n2} &= \sum_{i=1}^{n} \frac{R_i}{\pi_i(\hat{\eta})} \int_0^{x_i'\theta n^{-1/2}} [I(\epsilon_i \le s) - I(\epsilon_i \le 0)]\, ds \\ &= \sum_{i=1}^{n} \frac{R_i}{\pi_i(\eta_0)} \int_0^{x_i'\theta n^{-1/2}} [I(\epsilon_i \le s) - I(\epsilon_i \le 0)]\, ds \\ &+ \sum_{i=1}^{n} R_i \left(\frac{1}{\pi_i(\hat{\eta})} - \frac{1}{\pi_i(\eta_0)} \right) \int_0^{x_i'\theta n^{-1/2}} [I(\epsilon_i \le s) - I(\epsilon_i \le 0)] \\ &\equiv A_{n21} + A_{n22}. \end{aligned}$$

For A_{n21}:

$$\begin{aligned} A_{n21} &= E\left[A_{n21}|X\right] + A_{n21} - E\left[A_{n21}|X\right] \\ &= \sum_{i=1}^{n} \int_0^{x_i'\theta n^{-1/2}} F_i(s) - F_i(0) ds + o_p(1) \\ &= n^{-1}\theta' \sum_{i=1}^{n} f_i(0) x_i x_i' \theta + o_p(1) \xrightarrow{p} \Sigma_1. \end{aligned}$$

Then $A_{n2} \xrightarrow{p} \tilde{\Sigma}_1$ because for A_{n22}

$$\begin{aligned} A_{n22} &= (\hat{\eta} - \eta_0)' \sum_{i=1}^{n} R_i \left(\frac{\partial \pi_i(\eta)}{\partial \eta} \right)_{\eta=\eta_0} \frac{1}{\pi_i(\eta_0)^2} \int_0^{x_i'\theta n^{-1/2}} I(\epsilon_i \le s) - I(\epsilon_i \le 0) \\ &= (\hat{\eta} - \eta_0)' n^{-1} \sum_{i=1}^{n} \left(\frac{\partial \pi_i(\eta)}{\partial \eta} \right)_{\eta=\eta_0} \frac{1}{\pi_i(\eta_0)} f_i(0)(\theta' x_i)^2 = o_p(1). \end{aligned}$$

In Lemma 2 of Hjørt and Pollard (1993) it is shown that a minimizer of a convex function is asymptotically equivalent to the minimizer of a quadratic approximations of the convex function. Then by their basic corollary of Lemma 2

$$\hat{\theta} \xrightarrow{d} N(0, \tilde{\Sigma}^{-1}\tilde{\Sigma}_m\tilde{\Sigma}^{-1}).$$

□

Proof of Theorem 4.2

Proof is complete by demonstrating the following two steps:

1. Let β^* and $\tilde{\beta}$ be estimates from incorrect and correct models respectively. Then $\lim_{n\to\infty} \text{QBIC}_n(\beta^*) > \text{QBIC}_n(\tilde{\beta})$,

2. Let $\bar{\beta}$ and $\tilde{\beta}$ be estimates from correct models, but $\tilde{\beta}$ is the sparser model. Then $\lim_{n\to\infty} \text{QBIC}_n(\bar{\beta}) > \text{QBIC}_n(\tilde{\beta})$.

Let p^*, $\tilde{p}$ and $\bar{p}$ represent the number of parameters associated with β^*,$\tilde{\beta}$ and $\bar{\beta}$ respectively. Also, let $Z \sim N(0, \tilde{\Sigma}_m)$ Using Lemma 2 for some positive constant C

$$\begin{aligned}
& n^{-1}\left(\text{QBIC}_n(\beta^*) - \text{QBIC}_n(\tilde{\beta})\right) \\
= \; & n^{-1}\left(\text{QBIC}_n(\beta^*) - \text{QBIC}_n(\beta_0) - (\text{QBIC}_n(\tilde{\beta}) - \text{QBIC}_n(\beta_0))\right) \\
= \; & -n^{-1/2}(\beta^* - \beta_0)'Z + (\beta^* - \beta_0)'\tilde{\Sigma}_1(\beta^* - \beta_0) \\
+ \; & n^{-1/2}(\tilde{\beta} - \beta_0)'Z - (\tilde{\beta} - \beta_0)'\tilde{\Sigma}_1(\tilde{\beta} - \beta_0) + \frac{\log(n)(p^* - \tilde{p})}{2n} \\
\geq \; & C||\beta^* - \beta_0||^2.
\end{aligned}$$

Last inequality comes from $||\tilde{\beta} - \beta_0|| = O_p\left(n^{-1/2}\right)$ and $\tilde{\Sigma}_1$ is a positive definite matrix. Lower bound is positive by condition 9. For the second step

$$\begin{aligned}
& n^{-1}\left(\text{QBIC}_n(\bar{\beta}) - \text{QBIC}_n(\tilde{\beta})\right) \\
= \; & n^{-1}\left(\text{QBIC}_n(\bar{\beta}) - \text{QBIC}_n(\beta_0) - (\text{QBIC}_n(\tilde{\beta}) - \text{QBIC}_n(\beta_0))\right) \\
= \; & -n^{-1/2}(\bar{\beta} - \beta_0)'Z + (\bar{\beta} - \beta_0)'\tilde{\Sigma}_1(\bar{\beta} - \beta_0) \\
+ \; & n^{-1/2}(\tilde{\beta} - \beta_0)'Z - (\tilde{\beta} - \beta_0)'\tilde{\Sigma}_1(\tilde{\beta} - \beta_0) + \frac{\log(n)(\bar{p} - \tilde{p})}{2n}.
\end{aligned}$$

Since both $\tilde{\beta}$ and $\bar{\beta}$ are $\sqrt{n}$ consistent estimators the dominating term is $\frac{\log(n)(\bar{p}-\tilde{p})}{2n}$ which is positive for any n because $\bar{p} > \tilde{p}$.

Proof of Theorem 4.3

Proof: By convexity, Lemma 5 implies $||\hat{\beta}^W_{PL} - \beta_0|| = O_p\left(\sqrt{\frac{dJ_n}{n}}\right)$ thus proving the consistency of $\hat{\beta}^W_{PL}$. It also follows from Lemma 5 that $||W_B(\hat{\gamma} - \gamma_0)|| = O_p\left(\sqrt{dJ_n}\right)$. Using these facts and condition 4,

$$\begin{aligned} n^{-1}\sum_{i=1}^{n} f_i(0)\left(\hat{g}(z_i) - g_0(z_i)\right)^2 &= n^{-1}\sum_{i=1}^{n} f_i(0)\left(W(z_i)'(\hat{\gamma} - \gamma_0) - u_{ni}\right)^2 \\ &\leq n^{-1}\left(\hat{\gamma} - \gamma_0\right) W_B^2 \left(\hat{\gamma} - \gamma_0\right) + O_p\left(J_n^{-2r}\right) \\ &= O_p\left(n^{-1} dJ_n\right). \end{aligned}$$

Then by condition 1, $n^{-1}\sum_{i=1}^{n}\left(\hat{g}(z_i) - g_0(z_i)\right)^2 = O_p\left(n^{-1}dJ_n\right)$. □

Proof of Theorem 4.4

Proof: Let $\psi_\tau(\epsilon) = (\psi_\tau(\epsilon_1), ..., \psi_\tau(\epsilon_n))'$, $\hat{R} = \text{diag}\left(R_1\pi_1(\hat{\eta}), ..., R_n\pi_n(\hat{\eta})\right)$ and define

$$\tilde{\theta}_1 = \sqrt{n}\left(X^{*\prime} B_n X^*\right)^{-1} X^{*\prime}\hat{R}\psi_\tau(\epsilon).$$

Notice by Lemma 3

$$\begin{aligned} \tilde{\theta}_1 &= n^{-1/2}\left(\Sigma_1 + o_p(1)\right)^{-1}\Delta_n'\hat{R}\psi_\tau(\epsilon)(1 + o_p(1)) \\ &= \left(\Sigma_1 + o_p(1)\right)^{-1} n^{-1/2}\sum_{i=1}^{n}\frac{R_i}{\pi_i(\hat{\eta})}\delta_i\psi_\tau(\epsilon_i)(1 + o_p(1)). \end{aligned}$$

To verify asymptotic normality of $\tilde{\theta}_1$, we check the Lindeberg-Feller condition. Define $W_{ni} = \frac{R_i}{\pi_i(\tilde{\eta})}\delta_i\psi_\tau(\epsilon_i)$. For any $\omega > 0$ and using conditions 1, 2, 6, 7 and 8

$$\begin{aligned}
& \sum_{i=1}^{n} E\left[||W_{ni}||^2 I(||W_{ni}|| > \omega)\right] \\
\leq\ & \epsilon^{-2}\sum_{i=1}^{n} E||W_{ni}||^4 \\
\leq\ & C(n\epsilon)^{-2}\sum_{i=1}^{n} E\left(\psi_\tau^4(\epsilon_i)\left(\delta_i'\delta_i\right)^2\right)(1+o(1)) \\
\leq\ & Cn^{-2}\epsilon^{-2}\sum_{i=1}^{n} E(||\delta_i||^4) = O_p(1/n) = o_p(1).
\end{aligned}$$

Also by directly applying results from Theorem 4.1 we observe that

$$\frac{1}{n}\sum_{i=1}^{n} E(W_{ni}W_{ni}') \to \left(\Sigma_2^W - \Sigma_m\right).$$

Proof is complete because from Lemma 8 it follows that $\sqrt{n}(\hat{\beta}_{PL}^W - \beta_0) = \tilde{\theta}_1 + o_p(1)$. □

Proof of Theorem 4.5

Proof: Consider the unpenalized objective function

$$S_n(\beta,\gamma) = \frac{1}{n}\sum_{i=1}^{n}\rho_\tau(Y_i - x_i'\beta - W(z_i)'\gamma),$$

with subgradient $s(\beta,\gamma) = (s_0(\beta,\gamma), ..., s_p(\beta,\gamma), ..., s_{p+dJ_n}(\beta,\gamma))$ given by

$$\begin{aligned} s_j(\beta,\gamma) &= \frac{\tau}{n}\sum_{i=1}^{n} x_{ij} I(Y_i - x_i'\beta - W(z_i)'\gamma > 0) \\ &\quad + \frac{1-\tau}{n}\sum_{i=1}^{n} x_{ij} I(Y_i - x_i'\beta - W(z_i)'\gamma < 0) \\ &\quad - \frac{1}{n}\sum_{i=1}^{n} x_{ij} a_i \quad \text{for } 0 \le j \le p, \\ s_j(\beta,\gamma) &= \frac{\tau}{n}\sum_{i=1}^{n} W_j(z_i) I(Y_i - x_i'\beta - W(z_i)'\gamma > 0) \\ &\quad + \frac{1-\tau}{n}\sum_{i=1}^{n} W_j(z_i) I(Y_i - x_i'\beta - W(z_i)'\gamma < 0) \\ &\quad - \frac{1}{n}\sum_{i=1}^{n} W_j(z_i) a_i \quad \text{for } p+1 \le j \le p_n + dJ_n, \end{aligned}$$

where $a_i = 0$ if $Y_i - x_i'\beta - W(z_i)'\gamma \neq 0$, and $a_i \in [\tau - 1, \tau]$ otherwise. For ease of notation in this proof let $(\hat{\beta}, \hat{\gamma})$ represent the oracle estimator from (4.9). Following the proof of Theorem 5.3 it is sufficient to show that with probability approaching one

$$s_j\left(\hat{\beta},\hat{\gamma}\right) = 0, \ j = 0, 1, ..., q \text{ or } \ j = p+1, ..., p+dJ_n, \tag{4.10}$$

$$\left|\hat{\beta}_j\right| \ge (a + 1/2)\lambda, \ j = 1, ..., q, \tag{4.11}$$

$$\left|s_j\left(\hat{\beta},\hat{\gamma}\right)\right| \le \lambda, \ j = q+1, ..., p. \tag{4.12}$$

Convex optimization theory immediately provides (4.10) holds, while (4.11) holds from $\sqrt{n}$ consistency of $\hat{\beta}$ as stated in Theorem 4.4. Define $X_{Ai} \in \mathbb{R}^{q+1}$ as the vector of active linear covariates. Using the outline of the proof of Lemma 1 part (5.8) proof

will be complete if we show that

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n}\frac{R_i}{\pi_i(\hat{\eta})}x_{ij}\left[I(Y_i - {x_{Ai}}'\hat{\beta} - \hat{g}(z_i) \leq 0) - \tau\right]\right| > \lambda/(p-q)\right) \to 0 \ \ \forall j.$$

Using condition 7 with Taylor expansion and rate of λ

$$\frac{1}{n}\sum_{i=1}^{n}\frac{R_i}{\pi_i(\eta_0)}x_{ij}\left[I(Y_i - {x_{Ai}}'\hat{\beta} - \hat{g}(z_i) \leq 0) - \tau\right] + o(\lambda).$$

Notice

$$\text{Var}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{R_i}{\pi_i(\eta_0)}x_{ij}\left[I(Y_i - {x_{Ai}}'\hat{\beta} - \hat{g}(z_i) \leq 0) - \tau\right]\right) = O_p\left(n^{-1}\right).$$

For the expected value

$$\begin{aligned} & E\left[\frac{1}{n}\sum_{i=1}^{n}\frac{R_i}{\pi_i(\eta_0)}x_{ij}\left[I(Y_i - {x_{Ai}}'\hat{\beta} - \hat{g}(z_i) \leq 0) - \tau\right]\right] \\ = \ & E\left[\frac{1}{n}\sum_{i=1}^{n}x_{ij}f_i(0)(X_i'(\hat{\beta} - \beta_0) + \hat{g}(z_i) - g_0(z_i))\right]. \end{aligned}$$

Above expectation is goes to zero by Bounded Convergence Theorem. Proof is complete by rate of $n^{-1/2}\lambda^{-1} = o(1)$ stated in the conditions of the theorem. □

Chapter 5

Ultrahigh Dimensional Additive Partial Linear Regression

As high-dimensional data become common in diverse fields, tremendous efforts have recently been devoted to sparse regression problems. Most of the existing work have focused on estimating the conditional mean of the response variable. It is well recognized that high-dimensional data are often heterogeneous, for which focusing on the mean function alone may be misleading. One effective way of dealing with this complexity is to consider estimating conditional quantiles at different quantile levels, which not only provides a more complete picture of the conditional distribution, but also allows for a more realistic interpretation of sparsity. The latter point was particularly advocated in the recent work of Wang et al. (2012) and He et al. (2013), which allow different subsets of covariates to be relevant at different quantiles. An added advantage of the quantile regression framework is that it is naturally robust to heavy-tailed errors. This is especially beneficial for analyzing microarray data, which are often skewed even after the popular log transformation.

In this chapter we develop a flexible additive partial linear additive quantile regression model for analyzing high-dimensional data. Given a random sample

$\{Y_i, x_{i1}, ..., x_{ip_n}, z_{i1}, ..., z_{id}\}$, $i = 1, ..., n$, the model assumes

$$\begin{aligned} Y_i &= \beta_{00} + \beta_{01}x_{i1} + ... + \beta_{0(p_n)}x_{ip_n} + \sum_{k=1}^{d} g_{0k}(z_{ik}) + \epsilon_i \\ &= x_i'\beta_{\mathbf{0}} + g_0(z_i) + \epsilon_i, \end{aligned} \tag{5.1}$$

where $\beta_{\mathbf{0}}(\tau) = (\beta_{00}(\tau), \beta_{01}(\tau), ..., \beta_{0p_n}(\tau))'$ is a $p_n + 1$-dimensional vector of unknown parameters, $x_i = (1, x_{i1}, ..., x_{ip_n})'$, $z_i = (z_{i1}, ..., z_{id})'$, and $g_0(z_i) = \sum_{k=1}^{d} g_{0k}(z_{ik})$. The random errors satisfy $P(\epsilon_i \leq 0 \mid x_i, z_i) = \tau$ for some $0 < \tau < 1$. Hence, $x_i'\beta_{\mathbf{0}} + g_0(z_i)$ is the τth conditional quantile of Y_i given (x_i, z_i). For identifiability, we assume that $E[g_{0k}(z_{ik})] = 0$ $\forall k$. The difference in this model from those discussed in previous chapters is p_n increase with n. We are interested in the case p_n is of similar order or much larger than n. As an example, in microarray data analysis, the x_{ij}'s often correspond to the expression values of different genes, while the z_{ik}'s often correspond to one or more clinical variables, such as age, that have potential nonlinear effects. When p is fixed, semiparametric quantile regression model was considered by He and Shi (1996), He et al. (2002), Wang et al. (2009), among others.

We still approximate the nonparametric components using B-spline basis functions. First, we study the asymptotic theory of estimating the model (5.1) when p_n diverges. In our setting, this corresponds to the oracle model, i.e., the one we obtain if we know which covariates are important in advance. This is along the line of the work of Welsh (1989), Bai and Wu (1994) and He and Shao (2000) for M-regression with diverging number of parameters and possibly nonsmooth objective functions, which, however, were restricted to linear regression. Lam and Fan (2008) derived the asymptotic theory of profile kernel estimator for general semiparametric models with diverging number of parameter while assuming a smooth quasi-likelihood function. Second, we propose using a penalized regression estimator when p_n is of an exponential order of n and the model has a sparse structure. For the SCAD (Fan and Li,

2001) penalty, we derive the oracle property of the proposed estimator under relaxed conditions. It is also an interesting finding that solving the non-convex penalized estimator can be achieved via solving a series of quantile regression problems, which can be conveniently implemented using existing software packages.

Deriving the asymptotic properties of the penalized estimator is very challenging as we need to simultaneously deal with the nonsmooth loss function, non-convex penalty function, approximation of nonlinear functions and very high dimensionality. To deal with these challenges, we combine a recently developed convex-differencing method with the modern empirical process techniques. The convex-differencing method relies on a representation of the penalized loss function as the difference of two convex functions, which leads to a sufficient local optimality condition.(Wang et al., 2012) Empirical process techniques are introduced to derive various error bounds associated with the nonsmooth objective function which contains both high dimensional linear covariates and approximations of nonlinear components. It is worth pointing out that our approach is different from what was used in the recent literature for studying the theory of high-dimensional semiparametric mean regression and is able to considerably weaken the conditions required in the literature. In particular, we do not need moment conditions for the random error and allow it to depend on the covariates.

In the previous chapters we analyzed the fixed p setting and existing work on penalized semiparametric regression has been largely limited to this setting, see, for example, Bunea (2004), Liang and Li (2009), Wang and Xia (2009), Liu et al. (2011), Kai et al. (2011), Wang et al. (2011). Important progress in the high-dimensional p setting has been recently made by Xie and Huang (2009), still assumes $p < n$, for partial linear regression, Huang et al. (2010) for additive models, Li et al. (2011), $p = o(n)$, for semi-varying coefficient models, among others. Linear quantile regression with high-dimensional covariates was investigated by Belloni and Chernozhukov (2011) (LASSO penalty) and Wang et al. (2012) (non-convex penalty).

Tang et al. (2013) considered a two-step procedure for a nonparametric varying coefficients quantile regression model with a diverging number of nonparametric functional coefficients. They required two separate tuning parameters and quite complex design conditions.

In this chapter we present the additive partial linear additive quantile regression model in the high dimensional setting and discuss the properties of the oracle estimator. The oracle estimator differs from previous estimators we have considered because the size of the parametric component of the true model, q_n, can increase with the sample size. We then discuss the properties of a SCAD penalized objective function for a model with increasing parametric component and an increasing number of potential linear components allowing for $p_n >> n$. Our main theorem states that the penalized estimator retains the oracle property allowing for some exponential rates of growth of p_n in with relationship to n and any polynomial rate of growth. We assess the performance of our estimator via Monte Carlo simulations and apply our method to model birth weight while accounting for gene expression data. Theorems are presented at the end of the chapter with many of the details given in the Appendix.

5.1 Partially Linear Additive Quantile Regression Model with Diverging Number of Parameter

For high-dimensional inference, it is often assumed that the vector of coefficients $\beta_{\mathbf{0}}$ in model (5.1) is sparse, that is, most of its components are zero. Let $A = \{1 \leq j \leq p_n : \beta_{0j} \neq 0\}$ be the index set of nonzero coefficients and $q_n = |A|$ be the cardinality of A. Both A and q_n depend on τ, but for ease of notation τ is omitted. Without loss of generality, we assume that the first $q_n + 1$ components of $\beta_{\mathbf{0}}$ are non-zero and the remaining $p_n - q_n$ components are zero. Hence, we can write $\beta_{\mathbf{0}} = \left(\beta'_{01}, \mathbf{0}'_{p_n - q_n}\right)'$, where $\mathbf{0}_{p_n - q_n}$ denotes the $(p_n - q_n)$-vector of zeros. Let $X = (1_n, X_1, ..., X_{p_n})$ be

the $n \times (p_n + 1)$ matrix of linear covariates corresponding to the true underlying model, where 1_n is an n-vector of ones. Let $X_A = (1_n, X_1, ..., X_{q_n})$ be the submatrix consisting of the first $(q_n + 1)$ columns of X corresponding to the active covariates; and let $X_{A^c} = (X_{q_n+1}, ..., X_{p_n})$ be the submatrix consisting of the last $p_n - q_n$ columns of X. The row vectors of X_A and X_{A^c} are denoted as $x'_{A_1}, ..., x'_{A_n}$ and $x'_{A_1^c}, ..., x'_{A_n^c}$.

5.1.1 Oracle Estimator

We first study the estimator we would obtain when the index set A is known in advance, which we refer to as the oracle estimator. We allow q_n, the size of A, to increase with n which resonates with the perspective that a more complex model can be fitted when more data are collected. We continue to use B-splines to estimate the unknown non-linear functions and current notation is consistent with notation used in the previous chapters. The oracle estimator for $(\beta_0', \gamma_0')'$ is defined as $\left(\hat{\beta}', \hat{\gamma}'\right)'$, where $\hat{\beta} \equiv \left(\hat{\beta}_{\mathbf{1}}', \mathbf{0}'_{p_n - q_n}\right)'$ and

$$\left(\hat{\beta}_{\mathbf{1}}, \hat{\gamma}\right) = \underset{(\beta_{\mathbf{1}}, \gamma)}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \rho_\tau(Y_i - x_i' \left(\beta_{\mathbf{1}}', \mathbf{0}'_{p_n - q_n}\right)' - w(z_i)'\gamma). \tag{5.2}$$

To understand the properties of the oracle estimator we need to formally define the relationship between X_A and Z. A nuance we did not focus on in the previous chapter is that we only need to define a relationship between the active linear terms and the non-linear variables. Another change is that in this setting the number of predictors is not fixed.

We use a similar setup to the fixed dimension case outlined in Section 3.2. Let $X_A = \left[1_n \; X_{A_{(-1)}}\right]$ where 1_n is an n-dimensional vector of ones and $X_{A_{(-1)}} \in \mathbb{R}^{n \times q_n}$

with $X_{A_{(-1)}} = (X_1, ..., X_{q_n})$. Define the set $\mathcal{H}_r^d = \left\{ \sum_{k=1}^d h_k(z) \mid h_j \in \mathcal{H}_r \right\}$ and

$$\begin{aligned} h_j^*(\cdot) &= \arg\inf_{h_j \in \mathcal{H}_r^d} \sum_{i=1}^n E\left[f_i(0)(x_{ij} - h_j(z_i))^2\right] \ 1 \le j \le q_n, \\ h_0^*(\cdot) &= 0. \end{aligned}$$

Let x_{Aij} be the element of $X_{A(-1)}$ at the ith row and the jth column. Define $\delta_{ij} \equiv x_{Aij} - h_j^*(z_i)$ as the bias term from approximating x_{Aij} with an additive function of z_i. Let $\delta_i = \left(1, \delta_{i1}, ..., \delta_{i(q_n)}\right)' \in \mathbb{R}^{(q_n+1)}$, $i = 1, \ldots, n$, and $\Delta_n = (\delta_1, ..., \delta_n)' \in \mathbb{R}^{n\times(q_n+1)}$. Define H such that $H_{ij} = h_{j+1}^*(z_i)$ then $X_A = H + \Delta_n$. New conditions are required to handle that in this setting the number of columns of X and Δ_n can change with n.

Condition 11

(Conditions on the covariates) There exist a positive constant M_1 such that $|x_{ij}| \le M_1$, $\forall\ 1 \le i \le n$, $1 \le j \le p_n$ and $E[\delta_{ij}^4] \le M_2$, $\forall\ 1 \le i \le n$, $1 \le j \le q_n$. There exist finite positive constants c and C such that

$$c \le \lambda_{\max}\left(n^{-1} X_A X_A'\right) \le C, \quad c \le \lambda_{\max}\left(n^{-1}\Delta_n \Delta_n'\right) \le C. \qquad \square$$

Condition 12

(Condition on the true underlying model) There is an upper bound to the size of the oracle model. Specifically $q_n = O\left(n^{c_1}\right)$ for some $c_1 < \frac{1}{2}$. $\square$

Condition 11 guarantees that the asymptotic variance of X_A and Δ_n behave nicely, which allows for asymptotic analysis of $\hat{\beta}_{\mathbf{1}}$. Condition 12 ensures that the rate of growth of the oracle model is slow enough for good estimators. The following theorem summarizes the asymptotic properties of the oracle estimators.

Theorem 5.1

Assumes Conditions 1-5 and 11-12 hold. Then

$$\begin{aligned} ||\hat{\beta}_{\mathbf{1}} - \beta_{\mathbf{01}}|| &= O_p\left(\sqrt{n^{-1}q_n}\right), \\ n^{-1}\sum_{i=1}^{n}(\hat{g}(z_i) - g_0(z_i))^2 &= O_p\left(n^{-1}(q_n + dJ_n)\right). \end{aligned}$$

An interesting observation is that for the high dimensional case q_n is part of the rate of the estimation rate for $\hat{g}$. Xie and Huang (2009) established a similar rate for $\hat{\beta}_{\mathbf{1}}$ for a partial linear mean model (without the additive components), but we have a faster rate for estimating g_0.

Let $B_n = \text{diag}(f_1(0), \ldots, f_n(0))$, be an $n \times n$ diagonal matrix with entries of the conditional pdf of $\epsilon \mid x_i, z_i$ evaluated at zero. As q_n diverges, to investigate the asymptotic distribution of $\hat{\beta}_{\mathbf{1}}$, we consider estimating an arbitrary linear combination of the components of $\beta_{\mathbf{01}}$.

Theorem 5.2

Assume the conditions of Theorem 5.1 are satisfied. Let m be a finite positive integer and A_n be an $l \times (q_n + 1)$ matrix with l fixed and $A_n'A_n \to G$, a positive definite matrix, then

$$\sqrt{n}A_n\Sigma_n^{-1/2}\left(\hat{\beta}_{\mathbf{1}} - \beta_{\mathbf{01}}\right) \to N(0, G)$$

in distribution, where $\Sigma_n = T_n^{-1}S_nT_n^{-1}$ with $T_n = n^{-1}\Delta_n'B_n\Delta_n$ and $S_n = n^{-1}\tau(1-\tau)\Delta_n'\Delta_n$. □

Thus when considering fixed linear components the linear estimators are asymptotically normal. Interested in inducing sparsity, and ultimately defining an estimator with the oracle property, we minimize the following penalized objective function for

estimating (β_0, γ_0),

$$Q^P(\beta, \gamma) = \frac{1}{n}\sum_{i=1}^{n} \rho_\tau(Y_i - x_i'\beta - w(z_i)'\gamma) + \sum_{j=1}^{p_n} p_\lambda(|\beta_j|), \tag{5.3}$$

where $p_\lambda(\cdot)$ is a penalty function with tuning parameter λ. We restrict our attention to the popular SCAD and LASSO penalties.

5.1.2 Solving the Penalized Estimator

We propose a new and effective algorithm to solve the above penalized estimation problem. By observing that we can write $|\beta_j|$ as $\rho_\tau(\beta_j) + \rho_\tau(-\beta_j)$, then we recognize that the LASSO penalized objective function is equivalent to

$$\left(\hat{\beta}, \hat{\gamma}\right) = \underset{\beta\,\gamma}{\operatorname{argmin}} \sum_{i=1}^{n} \rho_\tau(Y_i - x_i'\beta - w(z_i)'\gamma) + \lambda \sum_{j=1}^{p_n} \rho_\tau(\beta_j) + \rho_\tau(-\beta_j). \tag{5.4}$$

The above minimization problem can be framed as an unpenalized quantile regression problem with $n + 2p_n$ augmented observations. We denote these augmented observations by $(Y_i^*, x_i^*, w(z_i)^*)$, $i = 1, \ldots, (n + 2p_n)$. The first n observations are those in the original data, that is $(Y_i^*, x_i^*, w(z_i)^*) = (Y_i, x_i, w(z_i))$, $i = 1, \ldots, n$; for the next p_n observations, we have $(Y_i^*, x_i^*, w(z_i)^*) = (0, \lambda e_{i-n}, 0)$, $i = n + 1, \ldots, n + p_n$; and the last p_n observations are given by $(Y_i^*, x_i^*, w(z_i)^*) = (0, -\lambda e_{i-n-p_n}, 0)$, $i = n + p_n + 1, \ldots, n + 2p_n$. Where e_j is a length p vector with a value of 1 at the jth position and zero otherwise. Thus for the LASSO penalty we have been able to frame (5.3) as an unpenalized quantile regression problem with $n^* = n+2p_n$, the augmented sample size, and $p^* = p_n + dJ_n$, the number of coefficients to estimate. With $n^* > p^*$ this problem can easily be solved using existing algorithms.

An important observation is that the SCAD penalized estimator can be obtained by iteratively solving unpenalized weighted quantile regression problems on a similar

set of augmented data. More specifically, applying the idea of the LLA algorithm (Zou and Li, 2008), we initialize by setting $\beta = 0$ and $\gamma = 0$ and use the approximation of $p_\lambda(|\beta|) \approx |\beta| p'_\lambda(|\beta|)$. Then for each step $t \geq 1$, we update the estimator by

$$\left(\hat{\beta}^t, \hat{\gamma}^t\right) = \underset{(\beta,\gamma)}{\operatorname{argmin}} \left\{ n^{-1} \sum_{i=1}^{n} \rho_\tau(Y_i - x_i'\beta - w(z_i)'\gamma) + \sum_{j=1}^{p_n} p'_\lambda\left(|\hat{\beta}_j^{t-1}|\right) |\beta_j| \right\}, \quad (5.5)$$

where $\hat{\beta}_j^{t-1}$ is the value of β_j at step $t-1$. Using the same notation we used to describe the algorithm for the LASSO penalty at step t of (5.5) can be solved using a similar augmented method. Now we have $(Y_i^*, x_i^*, w(z_i)^*) = (0, p'_\lambda\left(|\hat{\beta}_j^{t-1}|\right) e_{i-n}, 0)$, $i = n+1, \ldots, n+p_n$; and the last p_n observations are given by $(Y_i^*, x_i^*, w(z_i)^*) = (0, -p'_\lambda\left(|\hat{\beta}_j^{t-1}|\right) e_{i-n-p_n}, 0)$, $i = n+p_n+1, \ldots, n+2p_n$. In our simulations, we used the quantreg package in R and continue with the iterative procedure until $||\hat{\beta}^t - \hat{\beta}^{t-1}||_1 + ||\hat{\gamma}^t - \hat{\gamma}^{t-1}|| < 10^{-7}$.

5.1.3 Model Selection Theory

For model selection with increasing number of covariates we impose an additional condition on how quickly a signal can decay, which is needed to identify the underlying model.

Condition 13

(Condition on the signal) There exist positive constants c_2 and c_3 such that $2c_1 < c_2 < 1$ and $n^{(1-c_2)/2} \min_{1 \leq j \leq q_n} |\beta_{j0}| \geq c_3$. □

Due to the nonsmoothness and non-convexity of the penalized objective function $Q^P(\beta, \gamma)$, the classical KKT condition is not applicable to analyze the asymptotic properties of the penalized estimator. To investigate the asymptotic theory of the non-convex estimator for ultra-high dimensional partial linear additive quantile regression model, we explore the necessary condition for the local minimizer of a convex

differencing problem presented by (Tao and An, 1997). Wang et al. (2012) explored how to use these techniques for linear quantile regression with an increasing number of covariates. We extend it to the setting of additive partial linear quantile regression.

Our approach concerns with a non-convex objective function that can be expressed as the difference of two convex functions. Specifically, we consider objective functions belonging to the class

$$\mathbf{F} = \{f(x) : f(x) = k(x) - l(x), k(\cdot), l(\cdot) \text{ are both convex}\}.$$

This is a very general formulation that incorporates many different forms of penalized objective functions. The subdifferential of $k(x)$ at x_0 is defined as

$$\partial k(x_0) = \{t : k(x) \geq k(x_0) + (x - x_0)'t, \ \forall\ x\}.$$

Similarly, we can define the subdifferential of $l(x)$. Let $\text{dom}(k) = \{x : k(x) < \infty\}$ be the effective domain of k. A necessary condition for β^* to be a local minimizer of $F(\beta)$ is that β^* has a neighborhood U such that $\partial l(\beta) \cap \partial k(\beta^*) \neq \emptyset, \forall\, \beta \in U \cap \text{dom}(k)$ (see Lemma 16 in the Appendix).

To appeal to the above necessary condition for the convex differencing problem, notice that for the SCAD penalty $Q^P(\beta, \gamma)$ can be written as

$$Q^P(\beta, \gamma) = k(\beta, \gamma) - l(\beta),$$

where the two convex functions $k(\beta, \gamma) = n^{-1} \sum_{i=1}^{n} \rho_\tau(Y_i - x_i'\beta - w(z_i)'\gamma) + \lambda \sum_{j=1}^{p_n} |\beta_j|$, and $l(\beta) = \sum_{j=1}^{p_n} L(\beta_j)$ with,

$$\begin{aligned} L(\beta_j) &= \left[\left(\beta_j^2 + 2\lambda|\beta_j| + \lambda^2\right)/(2(a-1))\right] I(\lambda \leq |\beta_j| \leq a\lambda) \\ &+ \left[\lambda|\beta_j| - (a+1)\lambda^2/2\right] I(|\beta_j| > a\lambda). \end{aligned}$$

Building on the convex differencing structure, we show that with probability approaching one, the oracle estimator is a local minimizer of $Q^P(\beta, \gamma)$. To study the necessary optimality condition, we formally define $\partial k(\beta, \gamma)$ and $\partial l(\beta)$, the subdifferentials of $k(\beta)$ and $l(\beta)$, respectively. First, the function $l(\beta)$ does not depend on γ and is differentiable everywhere. Hence, its subdifferential is simply the regular derivative. For any value of β,

$$\partial l(\beta) = \left\{ \mu = (\mu_0, \mu_1, ..., \mu_{p_n})' \in \mathbb{R}^{p_n+1} : \mu_j = \frac{\partial l(\beta)}{\partial \beta_j} \right\}.$$

For the SCAD penalty function, $\frac{\partial l(\beta)}{\partial \beta_0} = 0$ and $\frac{\partial l(\beta)}{\partial \gamma_k} = 0$, $\forall\ k$. For $1 \leq j \leq p_n$,

$$\frac{\partial l(\beta)}{\partial \beta_j} = \begin{cases} 0, & 0 \leq |\beta_j| < \lambda, \\ (\beta_j - \lambda \text{sgn}(\beta_j))/(a-1), & \lambda \leq |\beta_j| \leq a\lambda, \\ \lambda \text{sgn}(\beta_j), & |\beta_j| > a\lambda. \end{cases}$$

On the other hand, the function $k(\beta, \gamma)$ is not differentiable everywhere. Its subdifferential at (β, γ) is a collection of vectors:

$$\begin{aligned}
\partial k(\beta,\gamma) &= \Bigg\{ \xi = (\xi_0, \xi_1, ..., \xi_{p_n}, \xi_{p_n+1}, ..., \xi_{p_n+dJ_n}) \in \mathbb{R}^{p_n+dJ_n+1} : \\
\xi_j &= -\tau n^{-1} \sum_{i=1}^{n} x_{ij} I(Y_i - x_i'\beta - w(z_i)'\gamma > 0) \\
&\quad + (1-\tau) n^{-1} \sum_{i=1}^{n} x_{ij} I(Y_i - x_i'\beta - w(z_i)'\gamma < 0) \\
&\quad - n^{-1} \sum_{i=1}^{n} x_{ij} a_i + \lambda l_j, \quad \text{for} \;\; 1 \le j \le p_n; \\
\xi_j &= -\tau n^{-1} \sum_{i=1}^{n} w_{j-p_n}(z_i) I(Y_i - x_i'\beta - w(z_i)'\gamma > 0) \\
&\quad + (1-\tau) n^{-1} \sum_{i=1}^{n} w_{j-p_n}(z_i) I(Y_i - x_i'\beta - w(z_i)'\gamma < 0) \\
&\quad - n^{-1} \sum_{i=1}^{n} w_{j-p_n}(z_i) a_i, \quad \text{for} \;\; p_n + 1 \le j \le p_n + dJ_n \Bigg\},
\end{aligned}$$

where $a_i = 0$ if $Y_i - x_i'\beta - w(z_i)'\gamma \neq 0$, and $a_i \in [\tau - 1, \tau]$ otherwise; $l_0 = 0$; for $1 \le j \le p_n$ $l_j = \text{sgn}(\beta_j)$ if $\beta_j \neq 0$ and $l_j \in [-1, 1]$ otherwise.

In the following, we analyze the subgradient of the unpenalized objective function, which plays an essential role in checking the optimality condition. The subgradient

$s(\beta,\gamma) = (s_0(\beta,\gamma), ..., s_{p_n}(\beta,\gamma), ..., s_{p_n+dJ_n}(\beta,\gamma))$ is given by

$$\begin{aligned} s_j(\beta,\gamma) &= \frac{\tau}{n}\sum_{i=1}^{n} x_{ij} I(Y_i - x_i'\beta - w(z_i)'\gamma > 0) \\ &\quad + \frac{1-\tau}{n}\sum_{i=1}^{n} x_{ij} I(Y_i - x_i'\beta - w(z_i)'\gamma < 0) \\ &\quad - \frac{1}{n}\sum_{i=1}^{n} x_{ij} a_i \quad \text{for } 0 \le j \le p_n, \\ s_j(\beta,\gamma) &= \frac{\tau}{n}\sum_{i=1}^{n} w_j(z_i) I(Y_i - x_i'\beta - w(z_i)'\gamma > 0) \\ &\quad + \frac{1-\tau}{n}\sum_{i=1}^{n} w_j(z_i) I(Y_i - x_i'\beta - w(z_i)'\gamma < 0) \\ &\quad - \frac{1}{n}\sum_{i=1}^{n} w_j(z_i) a_i \quad \text{for } p_n + 1 \le j \le p_n + dJ_n, \end{aligned}$$

where a_i is defined as before. The following lemma states the behavior of $s(\hat{\beta},\hat{\gamma})$ when being evaluated the oracle estimator.

Lemma 1

Assume Conditions 1-5 and 11-13 are satisfied and $\lambda = o\left(n^{-(1-c_2)/2}\right)$. For the oracle estimator $\left(\hat{\beta},\hat{\gamma}\right)$, with probability approaching one

$$\begin{aligned} s_j\left(\hat{\beta},\hat{\gamma}\right) &= 0, \ j = 0, 1, ..., q_n \text{ or } \ j = p_n + 1, ..., p_n + dJ_n, && (5.6) \\ \left|\hat{\beta}_j\right| &\ge (a + 1/2)\lambda, \ j = 1, ..., q_n, && (5.7) \\ \left|s_j\left(\hat{\beta},\hat{\gamma}\right)\right| &\le \lambda, \ j = q_n + 1, ..., p_n. && (5.8) \end{aligned}$$

□

Remark. Note that for $\xi_j \in \partial k(\beta, \gamma)$

$$\begin{aligned} \xi_0 &= s_j(\beta, \gamma), \\ \xi_j &= s_j(\beta, \gamma) + \lambda l_j, \quad \text{for } 1 \leq j \leq p_n, \quad l_j \in [-1, 1] \\ \xi_j &= s_j(\beta, \gamma), \quad \text{for } p_n + 1 \leq j \leq p_n + dJ_n. \end{aligned}$$

Thus Lemma 1 provides important insight on the asymptotic behavior of $\xi \in \partial k(\beta, \gamma)$.

Consider a small neighborhood around the oracle estimator $\left(\hat{\beta}, \hat{\gamma}\right)$ with radius $\lambda/2$. Building on Lemma 1, we prove in the Appendix that with probability tending to one, for any $(\beta, \gamma) \in \mathbb{R}^{p_n + dJ_n + 1}$ in this neighborhood, there exists $\xi = \left(\xi_0, ..., \xi_{p_n}, \mathbf{0}'_{dJ_n}\right)' \in \partial k(\beta, \gamma)$ such that

$$\frac{\partial l(\beta)}{\partial \beta_j} = \xi_j, \ j = 0, ..., p_n.$$

This leads to the main theorem of the paper. Let $\mathcal{E}_n(\lambda)$ be the set of local minima of $Q^P(\beta, \gamma)$. The theorem below shows that with probability approaching one, the oracle estimator belongs to the set $\mathcal{E}_n(\lambda)$.

Theorem 5.3

Assume conditions 1-5 and 11-13 are satisfied. Consider the SCAD penalty function with tuning parameter λ. Let $\hat{\eta} \equiv \left(\hat{\beta}, \hat{\gamma}\right)$ be the oracle estimator. If $\lambda = o\left(n^{-(1-c_2)/2}\right), n^{-1/2} q_n = o(\lambda)$, $n^{-1/2} J_n = o(\lambda)$ and $\log(p_n) = o(n\lambda^2)$, then

$$P\left(\hat{\eta} \in \mathcal{E}_n(\lambda)\right) \to 1$$

as $n \to \infty$. □

The above conditions for λ will hold for $\lambda = n^{-1/2+\delta}$ with $\delta \in (\max(c_1, 1/3), c_2/2)$.

Remark. The rates of λ are similar to those in Wang, Wu and Li (2012), but we

require the additional rate of $n^{-1/2}J_n = o(\lambda)$ to handle the additive partial linear setting.

Remark. The selection of the tuning parameter λ is important in practice. Cross-validation if known to often result in overfitting. Lee et al. (2013) recently proposed high-dimensional BIC for linear quantile regression when p is much larger than n. Motivated by their work, we consider the following high-dimensional BIC criterion.

$$\text{HQBIC}(\lambda) = \log\left(\sum_{i=1}^{n}\rho_\tau\left(Y_i - x_i'\hat{\beta}_\lambda - w(z_i)'_{J_n}\hat{\gamma}_\lambda\right)\right) + \nu_\lambda\frac{\log(p_n)\log(\log(n))}{2n}, \quad (5.9)$$

where p_n is the number of candidate linear covariates and $\nu(\lambda)$ is the number of degrees of freedom of the fitted model, which is the number of interpolated fits for quantile regression. We select the λ that minimizes the above criterion.

5.2 Simulation

In the Monte Carlo studies, we investigate the performance of the penalized additive partial linear quantile regression estimator in high dimension. We focus on the SCAD penalty and referred to the new procedure as Q-SCAD. The Q-SCAD is compared with three alternative procedures: additive partial linear quantile regression estimator with the LASSO penalty (Q-LASSO), additive partial linear mean regression with SCAD penalty (LS-SCAD) and LASSO penalty (LS-LASSO).

We first generate $\tilde{X} = (\tilde{X}_1, \ldots, \tilde{X}_{p+2})'$ from the multivariate normal distribution $N_{p+2}(0, \Sigma)$, where $\Sigma = (\sigma_{jk})_{(p+2)\times(p+2)}$ with $\sigma_{jk} = 0.5^{|j-k|}$. Then we set $X_1 = \sqrt{12}\Phi(\tilde{X}_1)$ where $\Phi(\cdot)$ is distribution function of $N(0,1)$ distribution and $\sqrt{12}$ scales X_1 to have standard deviation one. Furthermore, we let $Z_1 = \Phi(\tilde{X}_{25})$, $Z_2 = \Phi(\tilde{X}_{26})$,

$X_i = \tilde{X}_i$ for $i = 2, ..., 24$ and $X_i = \tilde{X}_{i-2}$ for $i = 27, ..., p+2$. The random responses are generated from the regression model

$$Y_i = x_{i6} + x_{i12} + x_{i15} + x_{i20} + \sin(2\pi z_{i1}) + z_{i2}^3 + \epsilon_i. \tag{5.10}$$

We consider three different distributions of the error term ϵ_i: (1) standard normal distribution; (2) t distribution with 3 degrees of freedom; and (3) heteroscedastic normal distribution with $\epsilon_i = \tilde{x}_{i1}\xi_i$ where $\xi_i \sim N(0,1)$ are independent of the X_i's.

To assess the performance of different methods, we use the following criteria:

1. False Variables (FV): average number of linear covariates incorrectly included in the model.

2. True Variables (TV): average number of linear covariates correctly included in the model.

3. True: proportion of times the true model is exactly identified.

4. P: proportion of times X_1 is selected.

5. AADE: average of the *average absolute deviation* (ADE) of the fit of the non-linear components, where the ADE is defined as $n^{-1}\sum_{i=1}^{n} |\hat{g}(z_i) - g_0(z_i)|$.

6. MSE: average of the mean squared error for estimating β_0, that is, average of $||\hat{\beta} - \beta_0||^2$.

The simulations have sample size $n = 300$ with $p = 100$, 300 and 600. We model $\tau = .5$ for error settings (1) and (2). For the heteroscedastic errors we model $\tau = .7$ and $\tau = .9$ and run 100 simulations for each setting. Note that at $\tau = 0.7$ or 0.9, when the error has the aforementioned heteroscedastic distribution, X_1 should also be included in the true model, that is, at these two quantiles the true model consists of 5 linear

covariates. In all simulations, the number of basis functions J_n is set to three, which we find to work satisfactorily in a variety of settings. For the LASSO method we select the tuning parameters λ by using five-fold cross validation. The simulation results are summarized in Table 5.1-Table 5.4. Table 5.1 and Table 5.2 report results for $\tau = 0.5$, with $N(0,1)$ and T_3 error distribution, respectively. Table 5.3 and Table 5.4 report results for the heteroscedastic error, $\tau = 0.7$ and 0.9, respectively. The least squares estimates of $\hat{\beta}$ for $\tau \neq .5$ are derived by assuming $\epsilon_i \sim N(0, \sigma)$. We note that the method with the SCAD penalty tends to pick a smaller and more accurate model. The advantages of quantile regression can be seen by the stronger performance for the quantile regression methods when the errors have a long tailed distribution such as T_3. Also, the quantile regression models do better at detecting the heteroscedastic terms. Our simulations indicate that it is harder to identify the heteroscedastic terms. Estimation of the non-linear terms is similar across error distributions and p.

The LASSO methods tend to select a larger model than the SCAD methods. The trade off between the two penalties is apparent in Table 5.3 where Q-LASSO correctly includes X_1 in the final model a higher percentage of the time than Q-SCAD. However, on average Q-LASSO includes a larger number of false variables than Q-SCAD. In Table 5.4 we see that both Q-SCAD and Q-LASSO select X_1 at a similar rate. This is because the signal of X_1 is stronger at $\tau = .9$. The LASSO procedure would be preferable to the practitioner if they are interested in detecting small signals. If the practitioner is worried about overfitting the model than the SCAD penalty would be better.

Method	n	p	FV	TV(4)	True	P	AADE	MSE
Q-SCAD	300	100	0.12	4.00	0.92	0.00	0.61	0.02
Q-LASSO	300	100	13.27	4.00	0.02	0.14	0.61	0.12
LS-SCAD	300	100	0.15	4.00	0.89	0.00	0.26	0.01
LS-LASSO	300	100	11.10	4.00	0.00	0.14	0.26	0.07
Q-SCAD	300	300	0.00	4.00	1.00	0.00	0.62	0.02
Q-LASSO	300	300	17.94	4.00	0.01	0.09	0.62	0.15
LS-SCAD	300	300	0.08	4.00	0.92	0.00	0.26	0.02
LS-LASSO	300	300	14.93	4.00	0.00	0.07	0.26	0.09
Q-SCAD	300	600	0.01	4.00	0.99	0.00	0.61	0.02
Q-LASSO	300	600	19.93	4.00	0.00	0.05	0.61	0.16
LS-SCAD	300	600	0.04	4.00	0.96	0.00	0.26	0.02
LS-LASSO	300	600	17.64	4.00	0.00	0.01	0.26	0.09

Table 5.1: High-dimensional simulation results for $\tau = .5$ and $\epsilon_i \sim N(0,1)$ Error N(0,1)

Method	n	p	FV	TV(4)	True	P	AADE	MSE
Q-SCAD	300	100	0.19	4.00	0.92	0.00	0.62	0.04
Q-LASSO	300	100	12.48	4.00	0.00	0.16	0.62	0.17
LS-SCAD	300	100	1.57	4.00	0.35	0.04	0.29	0.06
LS-LASSO	300	100	9.95	4.00	0.02	0.14	0.29	0.24
Q-SCAD	300	300	0.01	4.00	0.99	0.00	0.62	0.03
Q-LASSO	300	300	16.68	4.00	0.00	0.06	0.62	0.18
LS-SCAD	300	300	5.09	4.00	0.07	0.03	0.29	0.07
LS-LASSO	300	300	14.11	3.96	0.02	0.07	0.29	0.33
Q-SCAD	300	600	0.00	4.00	1.00	0.00	0.62	0.03
Q-LASSO	300	600	18.01	4.00	0.00	0.01	0.62	0.21
LS-SCAD	300	600	7.67	4.00	0.06	0.00	0.28	0.18
LS-LASSO	300	600	17.17	4.00	0.00	0.00	0.28	0.28

Table 5.2: High-dimensional simulation results $\tau = .5$ and $\epsilon_i \sim T_3$

Method	n	p	FV	TV(5)	True	P	AADE	MSE
Q-SCAD	300	100	0.26	4.87	0.71	0.87	0.62	0.04
Q-LASSO	300	100	13.37	4.98	0.00	0.98	0.62	0.15
LS-SCAD	300	100	0.71	4.01	0.01	0.01	0.27	0.71
LS-LASSO	300	100	11.24	4.19	0.00	0.19	0.27	0.81
Q-SCAD	300	300	0.15	4.69	0.59	0.69	0.61	0.07
Q-LASSO	300	300	18.26	4.90	0.00	0.90	0.61	0.21
LS-SCAD	300	300	1.36	4.00	0.00	0.00	0.27	0.69
LS-LASSO	300	300	15.79	4.11	0.00	0.11	0.27	0.81
Q-SCAD	300	600	0.19	4.64	0.47	0.64	0.62	0.08
Q-LASSO	300	600	20.20	4.89	0.00	0.89	0.62	0.23
LS-SCAD	300	600	1.87	4.02	0.00	0.02	0.27	0.67
LS-LASSO	300	600	17.39	4.09	0.00	0.09	0.27	0.80

Table 5.3: High-dimensional simulation results $\tau = .7$ and error Heteroscedastic

Method	n	p	FV	TV(5)	True	P	AADE	MSE
Q-SCAD	300	100	0.04	5.00	0.97	1.00	0.62	0.22
Q-LASSO	300	100	12.70	5.00	0.00	1.00	0.62	0.72
LS-SCAD	300	100	0.71	4.01	0.01	0.01	0.27	4.84
LS-LASSO	300	100	11.24	4.19	0.00	0.19	0.27	4.88
Q-SCAD	300	300	0.19	5.00	0.86	1.00	0.61	0.24
Q-LASSO	300	300	17.86	5.00	0.00	1.00	0.61	0.95
LS-SCAD	300	300	1.36	4.00	0.00	0.00	0.27	4.75
LS-LASSO	300	300	15.79	4.11	0.00	0.11	0.27	4.78
Q-SCAD	300	600	0.21	5.00	0.87	1.00	0.62	0.26
Q-LASSO	300	600	21.86	5.00	0.00	1.00	0.62	1.11
LS-SCAD	300	600	1.87	4.02	0.00	0.02	0.27	4.63
LS-LASSO	300	600	17.39	4.09	0.00	0.09	0.27	4.70

Table 5.4: High-dimensional simulation results $\tau = .9$ and error Heteroscedastic

5.3 Real Data Example

Votavova et al. (2011) obtained blood samples from peripheral blood, cord blood and the placenta from 20 pregnant smokers and 52 pregnant women without significant exposure to smoking. Their main objective was to identify differences in transcriptome alterations between the two groups. Birth weight of the baby (in kilograms) was recorded along with age of the mother, gestational age, parity, measurement of the amount of cotinine, a chemical found in tobacco, in the blood and mother's BMI. Low birth weight is known to be associated with both short-term and long-term health complications. Scientists are interested in which genes are associated with low birth weight.(Turan et al., 2012)

We consider modeling the 0.1, 0.3 and 0.5 conditional quantiles of infant birth weight. We use the gene data from the peripheral blood sample and have a total sample size of 64 after dropping samples with incomplete information. There are 24,526 expression values of probe sets. For preprocessing, we remove any probe sets for which the genes are not sufficiently expressed, that is, if the ratio between the maximum expression and minimum expression is less than 5. In addition, we removed any probe sets for which a single expression value is repeated 20 times or more. After these two preprocessing steps, 2,731 probe sets remain. For each quantile the top 200 probes are selected using the quantile-adaptive screening method proposed in He et al. (2013). The gene expression values of the 200 probes are included as linear covariates for the semiparametric quantile regression model. The clinical variables parity, gestational age, cotinine level and BMI. are also included as linear covariates. The effect of the age of the mother is modeled nonparametrically.

We consider the semiparametric quantile regression model with the SCAD and LASSO penalty functions. Least squares based semiparametric models with the SCAD and LASSO penalty functions are also considered. The tuning parameter

λ is selected by HQBIC for the SCAD estimator and by five-fold cross validation for LASSO as discussed in Section 4. The third column of Table 5.5 reports the number of nonzero elements, "Original NZ", for each model. As expected, the LASSO method selects a larger model than the SCAD penalty does. The number of non-zero variables varies with the quantile level, providing evidence that mean regression alone would provide a limited view of the conditional distribution.

τ	Method	Original NZ	Prediction Error	Randomized NZ
0.10	Q-SCAD	3	0.09 (0.05)	2.79 (2.78)
0.10	Q-LASSO	9	0.08 (0.02)	2.54 (3.04)
0.30	Q-SCAD	5	0.17 (0.04)	4.45 (4.45)
0.30	Q-LASSO	7	0.17 (0.03)	9.33 (8.97)
0.50	Q-SCAD	2	0.19 (0.05)	4.94 (3.41)
0.50	Q-LASSO	10	0.20 (0.04)	18.22 (11.9)
mean	LS-SCAD	3	0.19 (0.04)	2.92 (2.01)
mean	LS-LASSO	3	0.21 (0.04)	3.44 (2.81)

Table 5.5: Birth Weight Randomized Partition Results

Next, we compare different models on 100 random partitions of the data set. For each partition, we randomly select 50 subjects for the training data and 14 subjects for the test data. The fourth column of Table 5.5 reports the prediction error evaluated on the test data, defined as $14^{-1}\sum_{i=1}^{14}\rho_\tau(Y_i - \hat{Y}_i)$; while the fifth column reports the average number of linear covariates included in each model (denoted by "Randomized NZ"). Standard errors for both statistics are recorded in parentheses. We note that the SCAD method produces notably smaller models than the LASSO method without sacrificing much prediction accuracy.

We observe that different models for different random partitions. Table 5.6 summarizes the variables selected by Q-SCAD for $\tau = 0.1$, 0.3 and 0.5 and the frequency these variables are selected in the 100 random partitions. As expected gestational age has a strong signal across all quantiles. Probe ILMN_1656361 is another covariate found to have a very strong signal. The models for the 0.1 and 0.3 quantile are larger

on average.

Q-SCAD .1		Q-SCAD .3		Q-SCAD .5	
Covariate	Frequency	Covariate	Frequency	Covariate	Frequency
Gestational Age	93	Gestational Age	89	Gestational Age	96
ILMN_2279635	29	ILMN_1656361	39	ILMN_1656361	73
ILMN_1686871	3	ILMN_1696394	31		
		ILMN_1738921	20		
		ILMN_1714567	0		

Table 5.6: Variables selected by Q-SCAD in 100 random partitions

5.4 Proofs

Proof of Theorem 5.1

Proof for the rate of $n^{-1}\sum_{i=1}^{n}(\hat{g}(z_i) - g_0(z_i))^2$ given in Lemma 11. Proof for the rate of $||\hat{\beta} - \beta_0||$ follows from Lemmas 12 and 15.

Proof of Theorem 5.2

Proof follows from Lemmas 12 and 15.

Proof of part (5.6) of Lemma 1

Proof: Result follows from convex optimization theory. □

Proof of part (5.7) of Lemma 1

Proof: It is sufficient to show

$$P\left(\left|\hat{\beta}_j\right| \geq (a+1/2)\lambda, \text{ for } j = 1, ..., q_n\right) \to 1.$$

Notice that

$$\min_{1\leq j\leq q_n}\left|\hat{\beta}_j\right| \geq \min_{1\leq j\leq q_n}|\beta_{j0}| - \max_{1\leq j\leq q_n}\left|\hat{\beta}_j-\beta_{j0}\right|. \tag{5.11}$$

By condition 13 $\min_{1\leq j\leq q_n}|\beta_{j0}| \geq c_3 n^{-(1-c_2)/2}$, by theorem Theorem 5.1 $\max_{1\leq j\leq q_n}\left|\hat{\beta}_j-\beta_{j0}\right| = O_p\left(\sqrt{\frac{q_n}{n}}\right) = o_p\left(n^{-(1-c_2)/2}\right)$. Proof is complete by the assumption $\lambda = o\left(n^{-(1-c_2)/2}\right)$. □

Proof of Lemma 1 part (5.8)

Proof: Let $\mathcal{D} = \{i : Y_i - x'_{Ai}\hat{\beta} - W(z_i)'\hat{\gamma} = 0\}$, then for $j = q_n+1, ..., p_n$

$$s_j\left(\hat{\beta},\hat{\gamma}\right) = \frac{1}{n}\sum_{i=1}^{n} x_{ij}\left[I\left(Y_i - x'_{Ai}\hat{\beta} - W(z_i)'\hat{\gamma}\leq 0\right)-\tau\right] - \frac{1}{n}\sum_{i\in\mathcal{D}} x_{ij}(a_i^* + (1-\tau)),$$

where $a_i^* \in [\tau-1,\tau]$ when $i\in\mathcal{D}$ and for $j = 1, ..., q_n$ $s_j(\hat{\beta},\hat{\gamma}) = 0$ when $a_i = a_i^*$. Then with probability one $|\mathcal{D}| = d_n + 1$. Then

$$\frac{1}{n}\sum_{i\in\mathcal{D}} x_{ij}(a_i^* + (1-\tau)) = O_p\left(d_n n^{-1}\right) = o_p(\lambda).$$

Thus it is sufficient to show that

$$P\left(\max_{q_n+1\leq j\leq p_n}\left|\frac{1}{n}\sum_{i=1}^{n} x_{ij}\left[I(Y_i - x_{Ai}{}'\hat{\beta} - \hat{g}(z_i)\leq 0)-\tau\right]\right| > \lambda\right) \to 0.$$

Applying Lemmas 17 and 18

$$
\begin{aligned}
& P\left(\max_{q_n+1\le j\le p_n}\left|n^{-1}\sum_{i=1}^{n}x_{ij}\left[I(Y_i-x_{Ai}{}'\hat{\beta}-\hat{g}(z_i)\le 0)-\tau\right]\right|>\lambda\right)\\
\le\ & P\Bigg(\max_{q_n+1\le j\le p_n}\Bigg|n^{-1}\sum_{i=1}^{n}x_{ij}\Big[I(Y_i-x_{Ai}{}'\hat{\beta}-\hat{g}(z_i)\le 0)\\
& -I(Y_i-x_{Ai}{}'\beta_{\mathbf{01}}-g_0(z_i)\le 0)\Big]\Bigg|>\lambda/2\Bigg)\\
& +P\left(\max_{q_n+1\le j\le p_n}\left|n^{-1}\sum_{i=1}^{n}x_{ij}\left[I(Y_i-x_{Ai}{}'\beta_{\mathbf{01}}-g_0(z_i)\le 0)-\tau\right]\right|>\lambda/2\right)\\
\le\ & P\Bigg(\max_{q_n+1\le j\le p_n}\sup_{\substack{||\beta-\beta_{\mathbf{01}}||\le Cq_n^{1/2}n^{-1/2}\\ ||\gamma-\gamma_0||\le C\sqrt{\frac{d_n}{n}}}}\Bigg|n^{-1}\sum_{i=1}^{n}x_{ij}\Big[I(Y_i-x_{Ai}{}'\beta-W(z_i)'\gamma\le 0)\\
& -I(Y_i-x_{Ai}{}'\beta_{\mathbf{01}}-g_0(z_i)\le 0)\Big]\Bigg|>\lambda/2\Bigg)+o_p(1)\\
\le\ & P\Bigg(\max_{q_n+1\le j\le p_n}\sup_{\substack{||\beta-\beta_{\mathbf{01}}||\le Cq_n^{1/2}n^{-1/2}\\ ||\gamma-\gamma_0||\le C\sqrt{\frac{d_n}{n}}}}\Bigg|n^{-1}\sum_{i=1}^{n}x_{ij}\Big[I(Y_i-x_{Ai}{}'\beta-W(z_i)'\gamma\le 0)\\
& -I(Y_i-x_{Ai}{}'\beta_{\mathbf{01}}-g_0(z_i)\le 0)-P(Y_i-x_{Ai}{}'\beta-W(z_i)'\gamma\le 0)\\
& +P(Y_i-x_{Ai}{}'\beta_{\mathbf{01}}-g_0(z_i)\le 0)\Big]\Bigg|>\lambda/4\Bigg)\\
& +P\Bigg(\max_{q_n+1\le j\le p_n}\sup_{\substack{||\beta-\beta_{\mathbf{01}}||\le Cq_n^{1/2}n^{-1/2}\\ ||\gamma-\gamma_0||\le C\sqrt{\frac{d_n}{n}}}}\Bigg|n^{-1}\sum_{i=1}^{n}x_{ij}\Big[P(Y_i-x_{Ai}{}'\beta-W(z_i)'\gamma\le 0)\\
& -P(Y_i-x_{Ai}{}'\beta_{\mathbf{01}}-g_0(z_i)\le 0)\Big]\Bigg|>\lambda/4\Bigg)+o_p(1)\\
\le\ & P\Bigg(\max_{q_n+1\le j\le p_n}\sup_{\substack{||\beta-\beta_{\mathbf{01}}||\le Cq_n^{1/2}n^{-1/2}\\ ||\gamma-\gamma_0||\le C\sqrt{\frac{d_n}{n}}}}\Bigg|n^{-1}\sum_{i=1}^{n}x_{ij}\Big[P(Y_i-x_{Ai}{}'\beta-W(z_i)'\gamma\le 0)\\
& -P(Y_i-x_{Ai}{}'\beta_{\mathbf{01}}-g_0(z_i)\le 0)\Big]\Bigg|>\lambda/4\Bigg)+o_p(1).
\end{aligned}
$$

Notice

$$\begin{aligned}
& \max_{q_n+1\le j\le p_n} \sup_{\substack{||\beta-\beta_{\mathbf{01}}||\le Cq^{1/2}n^{-1/2} \\ ||\gamma-\gamma_0||\le C\sqrt{\frac{dJ_n}{n}}}} \left| n^{-1}\sum_{i=1}^{n} x_{ij}\Big[P(Y_i - x_{Ai}'\beta - W(z_i)'\gamma \le 0)\right. \\
& \left. -P(Y_i - x_{Ai}'\beta_{\mathbf{01}} - g_0(z_i) \le 0)\Big]\right| \\
= & \max_{q_n+1\le j\le p_n} \sup_{\substack{||\beta-\beta_{\mathbf{01}}||\le Cq^{1/2}n^{-1/2} \\ ||\gamma-\gamma_0||\le C_2\sqrt{\frac{dJ_n}{n}}}} \left| n^{-1}\sum_{i=1}^{n} x_{ij}\Big[F_i(x_{Ai}'(\beta_{\mathbf{1}} - \beta_{\mathbf{01}})\right. \\
& \left. +W(z_i)'(\gamma-\gamma_0) - u_{ni}) - F_i(0)\Big]\right| \\
\le & C \sup_{\substack{||\beta-\beta_{\mathbf{01}}||\le Cq^{1/2}n^{-1/2} \\ ||\gamma-\gamma_0||\le C\sqrt{\frac{d_n}{n}}}} n^{-1}\sum_{i=1}^{n} ||x_{Ai}||\cdot||\beta-\beta_{\mathbf{01}}|| + ||W(z_i)||\cdot||\gamma-\gamma_0|| + ||u_{ni}|| \\
\le & C\left(qn^{-1/2}\sqrt{dJ_n}\sqrt{d_n/n} + (dJ_n)^{-r}\right) = o(\lambda).
\end{aligned}$$

Note since $q_n n^{-1/2} = o(\lambda)$ and $J_n n^{-1/2} = o(\lambda)$ it follows that $\sqrt{dJ_n}\sqrt{d_n}n^{-1/2} = o(\lambda)$ and the proof is complete. □

Proof of Theorem 5.3

Proof: Recall that for $\xi_j \in \partial k(\beta,\gamma)$

$$\begin{aligned}
\xi_0 &= s_j(\beta,\gamma), \\
\xi_j &= s_j(\beta,\gamma) + \lambda l_j \text{ for } 1\le j\le p_n,\ \ l_j\in[-1,1] \\
\xi_j &= s_j(\beta,\gamma) \text{ for } p_n+1\le j\le p_n+dJ_n.
\end{aligned}$$

Define the set

$$\begin{aligned}\mathcal{G} = \Big\{ & \xi = (\xi_0, \xi_1, ..., \xi_{p_n}) : \xi_0 = 0; \xi_j = \lambda \text{sgn}(\hat{\beta}_j), j = 1, .., q_n \\ & \xi_j = s_j(\hat{\beta}, \hat{\gamma}) + \lambda l_j, j = q_n + 1, ..., p_n, \\ & \xi_j = 0, j = p_n + 1, ..., p_n + dJ_n. \Big\}\end{aligned}$$

and l_j ranges over [-1,1] for $j = q_n + 1, ..., p_n$. By Lemma 16 proof is complete if it is shown that there exists $\xi^* = \left(\xi_0^*, \xi_1^*, ..., \xi_{p_n}^*, ..., \xi_{p_n+dJ_n}\right)' \in \mathcal{G}$ in a neighborhood of $\lambda/2$ of $\left(\hat{\beta}, \hat{\gamma}\right)$ such that

$$P\left(\xi_j^* = \frac{\partial l(\beta)}{\partial \beta_j}, j = 0, ..., p_n + dJ_n\right) \to 1. \tag{5.12}$$

For $p_n + 1 \leq j \leq p_n + dJ_n$ $\frac{\partial l(\beta)}{\partial \beta_j} = 0$ and by Lemma 1

$$P\left(s_j(\hat{\beta}, \gamma) = 0\right) \to 1 \text{ for } j = p_n + 1, ..., p_n + dJ_n.$$

Therefore we only need to be concerned about the case of $0 \leq j \leq p_n$. In the following we define ξ_j^* so (5.12) is satisfied for $0 \leq j \leq p_n$.

1. For $j = 0$, $\xi_0^* = 0$ because $\frac{\partial l(\beta)}{\partial \beta_0} = 0$, it is immediate that $\frac{\partial l(\beta)}{\partial \beta_0} = \xi_0^*$.

2. For $j = 1, .., q_n$ $\xi_j^* = \lambda \text{sgn}(\hat{\beta}_j)$. For either penalty function $\frac{\partial l(\beta)}{\partial \beta_j} = \lambda \text{sgn}(\beta_j)$ for $|\beta_j| > a\lambda$. By Lemma 1 with probability one

$$\begin{aligned}\min_{1 \leq j \leq q_n} |\beta_j| &\geq \min_{1 \leq j \leq q_n} |\hat{\beta}_j| - \max_{1 \leq j \leq q_n} |\hat{\beta}_j - \beta_j| \\ &\geq (a + 1/2)\lambda - \lambda/2 = a\lambda.\end{aligned}$$

For any $1 \leq j \leq q_n$ $||\hat{\beta}_j - \beta_{j0}|| = O_p\left(n^{-1/2}\right) = o(\lambda)$ therefore for sufficiently

large n, $\hat{\beta}_j$ and β_j have the same sign.

3. By definition for $j = q_n+1, ..., p_n$ the oracle estimator has $\hat{\beta}_j = 0$ and $\left|\hat{\beta}_j - \beta_j\right| < \lambda$ therefore

$$|\beta_j| \leq |\hat{\beta}_j| + |\hat{\beta}_j - \beta_j| < \lambda.$$

For $\beta_j < \lambda$ then $\frac{\partial l(\beta)}{\partial \beta_j} = 0$ for the SCAD penalty. Therefore $P\left(\frac{l(\beta)}{\partial \beta_j} = 0\right) \to 1$. By Lemma 1 and the radius choice of $\lambda/2$

$$P\left(\frac{l(\beta)}{\partial \beta_j} \leq \lambda, j = q_n + 1, ..., p_n\right) \to 1.$$

By Lemma 1 $|s_j(\hat{\beta}_j)| \leq \lambda$ with probability approaching one for $j = q_n + 1, ..., p_n$. Therefore for both penalty functions there exists $l_j^* \in [-1, 1]$ such that $P(s_j(\hat{\beta}, \hat{\gamma}) + \lambda l_j^* = \frac{\partial l(\beta)}{\beta_j}, j = q_n + 1, ..., p_n) \to 1$. Define $\xi_j^* = s_j(\hat{\beta}, \hat{\gamma}) + \lambda l_j^*$.

From steps 1-3 above it follows that

$$P\left(\xi_j^* = \frac{\partial l(\beta)}{\partial \beta_j}, j = 0, ..., p_n\right) \to 1.$$

□

Chapter 6

Future Research

We have proposed quantile regression models to handle additive partial linear relationships, missing covariates and high-dimensional data. Our work in these areas leads to some natural extensions. We proposed using a weighted objective function to handle the bias caused by missing data. The weights were assigned by fitting a model to estimate the probability a subject would have complete data. If this method was misspecified then the estimator from the weighted objective function is no longer consistent. It would be helpful to have a procedure that is robust to misspecification of the weights. In mean regression Robins et al. (1994) proposed an augmented inverse probability weighting method. That is for the linear regression setting we solve

$$\sum_{i=1}^{n} \frac{R_i}{\pi(\hat{\eta})} x_i (Y_i - x_i'\hat{\beta}) + \left(1 - \frac{R_i}{\pi(\hat{\eta})}\right) \hat{m}(x_i, Y_i, \hat{\beta}) = 0,$$

where $\hat{m}(\cdot)$ is an estimate of $E\left[x_i(Y_i - x_i'\beta) \mid t_i\right]$, where t_i are the covariates that are always observed. Then the estimate of $\hat{\beta}$ is consistent if $\pi(\hat{\eta})$ or $\hat{m}(\cdot)$ are correctly specified making it a double robust procedure. For quantile regression it would be natural to consider

$$\sum_{i=1}^{n} \frac{R_i}{\pi(\hat{\eta})} x_i \psi_\tau (Y_i - x_i'\hat{\beta}) + \left(1 - \frac{R_i}{\pi(\hat{\eta})}\right) \hat{m}_\tau(x_i, Y_i, \hat{\beta}) = 0, \tag{6.1}$$

with $\hat{m}_\tau(\cdot)$ an estimate of $E\left[x_i\psi_\tau(Y_i - x_i'\beta) \mid t_i\right]$. However because the objective function of quantile regression is non-differentiable solving and $\psi_\tau(\cdot)$ is discrete we would need to consider an estimator that solves

$$\frac{1}{n}\sum_{i=1}^{n}\frac{R_i}{\pi(\hat{\eta})}x_i\psi_\tau(Y_i - x_i'\hat{\beta}) + \left(1 - \frac{R_i}{\pi(\hat{\eta})}\right)\hat{m}_\tau(x_i, Y_i, \hat{\beta}) = o_p(1). \tag{6.2}$$

How to solve (6.2) and understanding its asymptotic behavior remains an open research question.

In the additive partial linear models we limited model selection to the linear components of the model. For simultaneous model selection of the linear and non-linear terms we could consider a group penalty on the basis functions coefficients. Let γ_k be the basis coefficients corresponding to $w(z_{ik}) \in \mathbb{R}^{J_n}$ and consider the penalized objective function of

$$\sum_{i=1}^{n}\rho_\tau(Y_i - x_i'\beta - w(z_i)'\gamma) + \sum_{j=1}^{p_n}p_\lambda(|\beta_j|) + \sum_{k=1}^{d}p_\lambda(||\gamma_k||). \tag{6.3}$$

Finding (β, γ) which minimizes (6.3) could induce sparsity for both β and γ. The group penalty for the non-linear terms could send all J_n coefficients for a specific non-linear term to zero, implying that this variable has no relationship with the response. With the group penalty we could also consider estimation when the number of non-linear covariates increases with the sample size or use the group penalty to account for categorical predictors.

In our work we considered the error terms, ϵ_i, to be uncorrelated. This excludes a large number of data sets. For instance in Votavova et al. (2011) they used blood samples from peripheral blood, cord blood and the placenta, but to stay within the independent error framework we only considered measurements from the peripheral blood. He et al. (2002) proposed estimation of a partial linear model in the repeated

measurement setting, by showing that the estimates remain consistent when ignoring the correlation. That is say subject i has m_i observations and let Y_{ij}, x_{ij} and z_{ij} be the corresponding measurement for the ith subject at the jth observation. Then the following objective function is used

$$\sum_{i=1}^{n}\sum_{j=1}^{m_i}\rho_\tau(Y_{ij}-x_{ij}'\beta-w(z_{ij})'\gamma). \tag{6.4}$$

While Koenker (2004) proposed estimating an individual intercept, α_i, through a penalized objective function that incorporates data across τ. His objective function for the linear setting was

$$\sum_{k=1}^{K}\sum_{i=1}^{n}\sum_{j=1}^{m_i}W_k\rho_\tau(Y_{ij}-\alpha_i-x_{ij}'\beta(\tau_k))+\lambda\sum_{i=1}^{n}|\alpha_i|, \tag{6.5}$$

where W_k is a weight that controls the relative influence of the corresponding quantile τ_k. Showing that minimizing (6.4) produces consistent estimators for high-dimensional data would be a nice starting point. Generalizing (6.5) to handle high-dimensional and additive partial linear data would also be a nice results because (6.5) incorporates the repeated measurements of the data.

Quantile regression is a robust method that can model non-central behavior that would be ignored by ordinary least squares. We have presented models that create more flexible quantile regression models and allow for high-dimensional data, non-linear relationships or missing predictors. Future directions of research include making these procedures more robust, such as creating a method for missing data robust to misspecification of the missing pattern. Another direction would be to make these methods more general such as incorporating repeated measurement data. This dissertation outlined some steps to relax the linear quantile regression model, but there remains much work to be done.

References

Bai, Z. and Wu, Y. (1994). Limiting behavior of m-estimators of regression coefficients in high dimensional linear models i. scale-dependent case. *Journal of Multivariate Analysis*, 51:211–239.

Barrodale, I. and Roberts, F. (1974). Solution of an overdetermined system of equations in the l_1 norm. *Communications of the ACM*, 17:319–320.

Belloni, A. and Chernozhukov, V. (2011). L1-penalized quantile regression in high-dimensional sparse models. *Annals of Statistics*, 39:82–130.

Bunea, F. (2004). Consistent covariate selection and post model selection inference in semiparametric regression. *Annals of Statistics*, 32:898–927.

De Gooijer, J. and Zerom, D. (2003). On additive conditional quantiles with high-dimensional covariates. *Journal of the American Statistical Association*, 461:135–146.

Donald, S. and Newey, W. (1994). Series estimation of semilinear models. *Journal of Multivariate Analysis*, 50:30–40.

Eggleston, H. (1958). *Convexity*. Cambridge University Press.

Engel, E. (1857). Die productions- und consumtionsverhältnisse des königreichs sachsen. *Zeitschrift des statistischen Bureaus des Königlich Sächsischen Ministerium des Inneren*, 8:1–54.

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

He, X. and Shao, Q. (2000). On parameters of increasing dimensions. *Journal of Multivariate Analysis*, 73:120–135.

He, X. and Shi, P. (1994). Convergence rate of b-spline estimators of nonparametric conditional quantile functions. *Journal of Nonparametric Statistics*, 3:299–308.

He, X. and Shi, P. (1996). Bivariate tensor-product b-splines in a partly linear model. *Journal of Multivariate Analysis*, 58:162–181.

He, X., Wang, L., and Hong, H. (2013). Quantile-adaptive model-free nonlinear feature screening for high-dimensional heterogeneous data. *Annals of Statistics*, 41:342–369.

He, X., Zhu, Z., and Fung, W. (2002). Estimation in a semiparametric model for longitudinal data with unspecified dependence structure. *Biometrika*, 89:579–590.

Hjørt, N. and Pollard, D. (1993). Asymptotics for minimisers of convex processes. *Statistical Research Report*.

Horowitz, J. and Lee, S. (2005). Nonparametric estimation of an additive quantile regression model. *Journal of the American Statistical Association*, 472:1238–1249.

Huang, J., Horowitz, J., and Wei, F. (2010). Variable selection in nonparametric additive models. *Annals of Statistics*, 38:2282–2313.

Kai, B., Li, R., and Zou, H. (2011). New efficient estimation and variable selection methods for semiparametric varying-coefficient partially linear models. *Annals of Statistics*, 39:305–332.

Knight, K. (1998). Limiting distributions for l_1 regression estimators under general conditions. *Annals of Statistics*, 26:755–770.

Koenker, R. (2004). Quantile regression for longitudinal data. *Journal of Multivariate Analysis*, 91:74–89.

Koenker, R. (2005). *Quantile Regression*. Cambridge University Press.

Koenker, R. (2012). Pacakge reference manual: quantreg.

Koenker, R. and Bassett, G. (1978). Regression quantiles. *Econometrica*, 46:33–50.

Koenker, R. and Bassett, G. (1982). Robust tests of heteroscedasticity based on regression quantiles. *Econometrica*, 50:43–61.

Koenker, R. and d'Orey, V. (1987). Computing regression quantiles. *Applied Statistics*, 36:383–393.

Lam, C. and Fan, J. (2008). Profile-kernel likelihood inference with diverging number of parameters. *Annals of Statistics*, 36:2232–2260.

Lee, E., Noh, H., and Park, B. (2013). Model selection via bayesian information criterion for quantile regression models. *To appear in Journal of the American Statistical Association*.

Li, G., Xue, L., and Lian, H. (2011). Semi-varying coefficient models with a diverging number of components. *Journal of Multivariate Analysis*, 102:1166–1174.

Liang, H. and Li, R. (2009). Variable selection for partially linear models with measurement errors. *Journal of the American Statistical Association*, 104(485):234–248.

Liang, H., Wang, S., Robins, J., and Caroll, R. (2004). Estimation in partially linear models with missing covariates. *Journal of the American Statistical Association*, 99:357–367.

Lipsitz, S., Fitzmaurice, G., Molenberghs, G., and Zhao, L. (1997). Quantile regression methods for longitudinal data with drop-outs: application to cd4 cell counts of patients infected with the human immunodeficiency virus. *Journal of the Royal Statistical Society: Series C*, 46:463–476.

Little, R. and Rubin, D. (2002). *Statistical analysis with missing data: Second Edition*. Wiley.

Liu, X., Wang, L., and Liang, H. (2011). Estimation and variable selection for semiparametric additive partial linear models. *Statistica Sinica*, 21:1225–1248.

Machado, J. (1993). Robust model selection and m-estimation. *Econometric Theory*, 9:478–493.

Robins, J., Rotnitsky, A., and Zhao, L. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89:846–866.

Schumaker, L. (1981). *Spline Functions: Basic Theory*. Wiley: New York.

Sherwood, B., Wang, L., and Zhou, X. (2013). Weighted quantile regression for analyzing health care cost data with missing covariates. *Statistics in Medicine*, 32:4967–4979.

Shi, P. and Li, G. (1995). Global convergence rates of b-spline m-estimators in nonparametric regression. *Statistica Sinica*, 5:303–318.

Stone, C. (1982). Optimal global rates of convergence for nonparametric regression. *Annals of Statistics*, 10:1040–1053.

Stone, C. (1985). Additive regression and other nonparametric models. *Annals of Statistics*, 13:689–706.

Tang, Y., Song, X., Wang, H., and Zhu, Z. (2013). Variable selection in high-dimensional quantile varying coefficient models. *Journal of Multivariate Analysis*, 122:115–132.

Tao, P. and An, L. (1997). Convex analysis approach to d.c. programming: theory, algorithms and applications. *Acta Mathematica Vietnamica*, 22:289–355.

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

Tierney, W., Fitzgerald, J., Miller, M., Zhou, X., Harris, L., and Wolinsky, F. (1995). Predicting inpatient costs with admitting clinical data. *Medical Care*, 33:1–14.

Tsiatis, A. (2006). *Semiparametric Theory and Missing Data*. Springer.

Turan, N., Ghalwash, M., Kataril, S., Coutifaris, C., Obradovic, Z., and Sapienza1, C. (2012). Dna methylation differences at growth related genes correlate with birth weight: a molecular signature linked to developmental origins of adult disease? *BMC Medical Genomics*, 5:10.

Votavova, H., Dostalova Merkerova, M., Fejglova, K., Vasikova, A., Krejcik, Z., Pastorkova, A., Tabashidze, N., Topinka, J., Veleminsky, M. J., Sram, R., and Brdicka, R. (2011). Transcriptome alterations in maternal and fetal cells induced by tobacco smoke. *Placenta*, 32:763–770.

Wang, H. and Xia, Y. (2009). Shrinkage estimation of the varying coefficient model. *Journal of the American Statistical Association*, 486:747–757.

Wang, H., Zhu, Z., and Zhou, J. (2009). Quantile regression in partially linear varying coefficient models. *Annals of Statistics*, 37:3841–3866.

Wang, L. (2009). Wilcoxon-type generalized bayesian information criterion. *Biometrika*, 96:163–173.

Wang, L., Liu, X., Liang, H., and Carroll, R. (2011). Estimation and variable selection for generalized additive partial linear models. *Annals of Statistics*, 39:1827–1851.

Wang, L., Wu, Y., and Li, R. (2012). Quantile regression for analyzing heterogeneity in ultra-high dimension. *Journal of the American Statistical Association*, 107:214–22.

Wei, Y., Ma., Y., and Carroll, R. (2012). Multiple imputation in quantile regression. *Biometrika*, 99:423–438.

Welsh, A. (1989). On m-processes and m-estimation. *Annals of Statistics*, 17:337–361.

Wu, Y. and Liu, Y. (2009). Variable selection in quantile regression. *Statistica Sinica*, 19:801–817.

Xie, H. and Huang, J. (2009). Scad-penalized regression in high-dimensional partially linear models. *Annals of Statistics*, 37(2):673–696.

Yi, G. and He, W. (2009). Median regression models for longitudinal data with dropouts. *Biometrics*, 65:618–625.

Zhou, S., Shen, X., and Wolfe, D. (1998). Local asymptotics for regression splines and confidence regions. *The Annals of Statistics*, 26:1760–1782.

Zhou, X., K., S., and Tierney, W. (2001). Regression analysis of health care charges with heteroscedasticity. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 50:303–312.

Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Annals of Statistics*, 36:1509–1533.

Chapter 7

Appendix

Throughout the appendix, we use C to denote a positive constant which does not depend on n and may vary from line to line.

7.1 Lemmas for Chapter 4

7.1.1 Definitions

First we introduce, or recall, the following definitions

$$\begin{aligned}
B &= \text{diag}(f_1(0), \ldots, f_n(0)) \in \mathbb{R}^{n\times n}, \\
W &= (W(z_1), ..., W(z_n))' \in \mathbb{R}^{n\times dJ_n}, \\
P_W(B) &= W(W'BW)^{-1}W'B \in \mathbb{R}^{n\times n}, \\
X^* &= (x_1^*, \ldots, x_n^*)' \\
&= (1, X_1^*, ..., X_p^*) \in \mathbb{R}^{n\times p+1}, \\
W_B^2 &= W'BW \in \mathbb{R}^{dJ_n\times dJ_n}, \\
\Delta_n &= [1_n,\ \Delta_{n1},\ ..., \Delta_{np}] \in \mathbb{R}^{n\times p+1} \\
\Delta_n^B &= \Delta_n' B \Delta_n \in \mathbb{R}^{(p+1)\times(p+1)}, \\
\theta_1 &= \sqrt{n}\,(\beta - \beta_0) \in \mathbb{R}^{p+1}, \\
\theta_2 &= W_B\,(\gamma - \gamma_0) + W_B^{-1}W'BX(\beta - \beta_0) \in \mathbb{R}^{dJ_n}, \\
\tilde{x}_i &= n^{-1/2}x_i^* \in \mathbb{R}^{p+1}, \\
\tilde{W}(z_i) &= W_B^{-1}W(z_i) \in \mathbb{R}^{dJ_n}, \\
\tilde{s}_i &= \left(\tilde{x}_i', \tilde{W}(z_i)\right)' \in \mathbb{R}^{p+dJ_n+1}, \\
b_n &= dJ_n.
\end{aligned}$$

Let a_n be a sequence of positive numbers. Define

$$\begin{aligned}
Q_i(a_n) \equiv Q_i(a_n\theta_1, a_n\theta_2) &= \rho_\tau\left(\epsilon_i - a_n\tilde{x}_i'\theta_1 - a_n\tilde{W}(z_i)'\theta_2 - u_{ni}\right), \\
E_s[Q_i] &= E\left[Q_i \mid x_i, z_i\right].
\end{aligned}$$

It is noted that

$$n^{-1}\sum_{i=1}^{n}\rho_\tau(Y_i - x_i'\beta - W(z_i)'\gamma) = n^{-1}\sum_{i=1}^{n}\rho_\tau(\epsilon_i - \tilde{x}_i'\theta_1 - \tilde{W}(z_i)'\theta_2 - u_{ni}).$$

Define

$$\left(\hat{\theta}_1, \hat{\theta}_2\right) = \arg\min_{(\theta_1,\theta_2)} n^{-1}\sum_{i=1}^{n}\rho_\tau(\epsilon_i - \tilde{x}_i'\theta_1 - \tilde{W}(z_i)'\theta_2 - u_{ni}).$$

Define

$$D_i(\theta, a_n) = Q_i(a_n) - Q_i(0) - E_s\left[Q_i(a_n) - Q_i(0)\right] + a_n\left(\tilde{x}_i'\theta_1 + \tilde{W}(z_i)'\theta_2\right)\psi_\tau(\epsilon_i). \quad (7.1)$$

Noting that $\rho_\tau(u) = \frac{1}{2}|u| + \left(\tau - \frac{1}{2}\right)u$, we have

$$\begin{aligned} Q_i(a_n) - Q_i(0) &= \frac{1}{2}\left[\left|\epsilon_i - a_n\tilde{x}_i'\theta_1 - a_n\tilde{W}(z_i)'\theta_2 - u_{ni}\right| - |\epsilon_i - u_{ni}|\right] \\ &\quad - \left(\tau - \frac{1}{2}\right)\left(\tilde{x}_i'\theta_1 a_n + \tilde{W}(z_i)'\theta_2 a_n\right). \end{aligned} \quad (7.2)$$

Define

$$Q_i^*(a_n) = \frac{1}{2}\left[\left|\epsilon_i - \tilde{x}_i'\theta_1 a_n - \tilde{W}(z_i)'\theta_2 a_n - u_{ni}\right| - |\epsilon_i - u_{ni}|\right],$$

then by combining (7.1) and (7.2),

$$D_i(\theta, a_n) = Q_i^*(a_n) - E_s[Q_i^*(a_n)] + a_n\left(\tilde{x}_i'\theta_1 + \tilde{W}(z_i)'\theta_2\right)\psi_\tau(\epsilon_i). \quad (7.3)$$

The above simplification will be used in lemma 4. First we need to establish that x_i^* is an approximation of δ_i, which is important for understanding the asymptotic behavior of the additive partial linear estimators.

7.1.2 Rates for Basis functions

By Stone (1985) $||W(z_i)|| = O_p(1)$, $\forall\ i$. Applying the properties of the spline basis functions given in Zhou et al. (1998), it is immediate that $||W_B^{-1}|| = O_p\left(\sqrt{\frac{b_n}{n}}\right)$. Following Lemma 5.1 of Shi and Li (1995), it can be shown that with probability one $\max_i ||\tilde{W}(z_i)|| \leq C_0\sqrt{\frac{b_n}{n}}$, for some positive constant C_0. By the definition of $\tilde{x}_i$ and Condition 2, $\max_i ||\tilde{x}_i|| \leq C_1 n^{-1/2}$, for some positive constant C_1. By the result of Schumaker (1981), there exists a positive constant C_3, such that $\sup_{t\in[0,1]} |u_{ni}| \leq C_3 J_n^{-r}$.

7.1.3 Lemmas for Theorem 4.2

Lemma 2

If the conditions of theorem Theorem 4.2 hold then for $Z \sim N(0, \tilde{\Sigma}_m)$

$$\begin{aligned} & n^{-1}\sum_{i=1}^{n}\frac{R_i}{\pi_i(\hat{\eta})}\left(\rho_\tau(Y_i - x_i'\beta) - \rho_\tau(Y_i - x_i'\beta_0)\right) \\ = \ & -n^{-1/2}(\beta-\beta_0)'Z + (\beta-\beta_0)'\tilde{\Sigma}_1(\beta-\beta_0) + o_p(1). \end{aligned}$$

Proof: By Knights Identity (Knight, 1998)

$$\begin{aligned} n^{-1}\left(\text{QBIC}_n(\beta) - \text{QBIC}(\beta_0)\right) &= -n^{=1}(\beta-\beta_0)'\sum_{i=1}^{n} x_i\psi_\tau(\epsilon_i) \\ &+ n^{-1}\sum_{i=1}^{n}\int_0^{x_i'(\beta-\beta_0)} I(\epsilon_i \leq s) - I(\epsilon_i \leq 0)ds. \end{aligned}$$

Following results from proof of Theorem 4.1 we get:

1. $n^{-1/2}\sum_{i=1}^{n}\frac{R_i}{\pi_i(\hat{\eta})}x_i\psi_\tau(\epsilon_i) \xrightarrow{d} Z$,
2. $n^{-1}\sum_{i=1}^{n}\frac{R_i}{\pi_i(\hat{\eta})}\int_0^{x_i'(\beta-\beta_0)} I(\epsilon_i \leq s) - I(\epsilon_i \leq 0)ds \xrightarrow{p} (\beta-\beta_0)'\tilde{\Sigma}_1(\beta-\beta_0)$.

□

7.1.4 Lemmas for Theorem 4.3

Lemma 3

If conditions 3 - 5 hold then

(1) $n^{-1/2}X^* = n^{-1/2}\Delta_n + o_p(1)$,

(2) $n^{-1}X^{*\prime}B_nX^* = \Sigma_1 + o_p(1)$. □

Proof: Let $\Delta_{n(-1)} = [\Delta_{n1}, ..., \Delta_{np}]$. Then by the definition of X^* and Δ_n sufficient to show ,

$$n^{-1/2}(X_{(-1)} - P_W(B)X_{(-1)}) = n^{-1/2}\Delta_{n(-1)}.$$

Notice

$$n^{-1/2}(X_{(-1)} - P_W(B)X_{(-1)}) = n^{-1/2}\Delta_{n(-1)} + n^{-1/2}(H - P_W(B)X_{(-1)}).$$

Then consider the following weighted least squares problem. Let $\gamma_j^* \in \mathbb{R}^{dJ_n}$ be defined as $\gamma_j^* = \underset{\gamma\in\mathbb{R}^{dJ_n}}{\text{argmin}} \sum_{i=1}^n f_i(0)(x_{ij} - W(z_i)'\gamma)^2$. Let $\hat{h}_j(z_i) = W(z_i)'\gamma_j^*$ and notice that $\{P_W(B)X_{(-1)}\}_{ij} = \hat{h}_j(z_i)$. Adapting the results from Stone (1985), it follows that

$$\begin{aligned} n^{-1}||H - P_W(B)X_{(-1)}||^2 &= n^{-1}\lambda_{\max}\left((H - P_W(B)X_{(-1)})'(H - P_W(B)X_{(-1)})\right) \\ &\leq n^{-1}\text{trace}\left[(H - P_W(B)X_{(-1)})'(H - P_W(B)X_{(-1)})\right] \\ &= n^{-1}\sum_{i=1}^n\sum_{j=1}^p(h_j^*(z_i) - \hat{h}_j(z_i))^2 \\ &= O_p\left(\frac{J_n}{n}\right) = o_p(1), \end{aligned}$$

by conditions 3 and 5. The lemma follows immediately. □

Lemma 4

If the conditions of Theorem 4.3 hold then for any $\omega > 0$

$$P\left(\sup_{\substack{||\theta|| \leq L \\ ||\eta-\eta_0|| \leq Cn^{-1/2}}} b_n^{-1}\left|\sum_{i=1}^{n}\frac{R_i}{\pi_i(\eta)}D_i(\theta,\sqrt{b_n})\right|\right) > \omega. \qquad \square$$

Proof: Let A_{n1} denote the event $\tilde{s}_{(n)} \leq C_2\sqrt{b_n/n}$, where $\tilde{s}_{(n)} = \max_i ||\tilde{s}_i||$ and $C_2 = \max(C_0, C_1)$. Let A_{n2} denote the event $\max_i |u_{ni}| \leq C_3 J_n^{-r}$. Then by rates discussed in Section 7.1.2 $P(A_{n1}) = 1$ and $P(A_{n2}) = 1$. To prove the lemma, it is sufficient to show that $\forall\ \omega > 0$,

$$P\left(\sup_{\substack{||\theta|| \leq 1 \\ ||\eta-\eta_0|| \leq Cn^{-1/2}}} \left|b_n^{-1}\sum_{i=1}^{n}\frac{R_i}{\pi_i(\eta)}D_i\left(\theta, L\sqrt{b_n}\right)\right| > \omega, A_{n1}\cap A_{n2}\right) \to 0. \tag{7.4}$$

Note that

$$\begin{aligned}
& P\left(\sup_{\substack{||\theta|| \leq 1 \\ ||\eta-\eta_0|| \leq Cn^{-1/2}}} \left|b_n^{-1}\sum_{i=1}^{n}\frac{R_i}{\pi_i(\eta)}D_i\left(\theta, L\sqrt{b_n}\right)\right| > \omega, A_{n1}\cap A_{n2}\right) \\
\leq\ & P\left(\sup_{||\theta|| \leq 1} \left|b_n^{-1}\sum_{i=1}^{n}\frac{R_i}{\pi_i(\eta_0)}D_i\left(\theta, L\sqrt{b_n}\right)\right| > \omega/2, A_{n1}\cap A_{n2}\right) \\
+\ & P\left(\sup_{\substack{||\theta|| \leq 1 \\ ||\eta-\eta_0|| \leq Cn^{-1/2}}} \left|b_n^{-1}\sum_{i=1}^{n}R_i\left(\frac{1}{\pi_i(\eta)}-\frac{1}{\pi_i(\eta_0)}\right)D_i\left(\theta, L\sqrt{b_n}\right)\right| > \omega/2, A_{n1}\cap A_{n2}\right) \\
\equiv\ & P_{n1} + P_{n2}.
\end{aligned}$$

Where P_{n1} and P_{n2} are probability statements whose definitions follow directly from the above statement. First we will show that $\lim_{n\to\infty} P_{n1} = 0$. Define

$$\Theta \equiv \left\{\theta \mid ||\theta|| \leq 1, \theta \in \mathbb{R}^{b_n+1}\right\},$$

we can partition Θ as a union of disjoint regions $\Theta_1, ..., \Theta_{M_n}$, such that the diameter of each region does not exceed $m_0 = \frac{\omega\alpha_l}{8C_2L\sqrt{n}}$, where α_l is defined in condition 6. Then for some positive constant C a covering can be constructed such that $M_n \leq C\left(\frac{C\sqrt{n}}{\omega}\right)^{b_n+1}$. Let $\theta_1^*, ..., \theta_{M_n}^*$ be arbitrary points in $\Theta_1, ..., \Theta_{M_n}$, respectively, and write $\theta_k^* = (\theta_{k1}^{*\prime}, \theta_{k2}^{*\prime})'$, $k = 1, \ldots, M_n$.

Then

$$\begin{aligned}
& P\left(\sup_{||\theta|\ \leq 1} b_n^{-1}\left|\sum_{i=1}^{n}\frac{R_i}{\pi_i(\eta_0)}D_i(\theta, L\sqrt{b_n})\right| > \omega/2, A_{n1}\cap A_{n2}\right) \\
\leq & \sum_{k=1}^{M_n} P\left(\sup_{||\theta|\ \leq 1} b_n^{-1}\left|\sum_{i=1}^{n}\frac{R_i}{\pi_i(\eta_0)}D_i(\theta, L\sqrt{b_n})\right| > \omega/2, A_{n1}\cap A_{n2}\right) \\
\leq & \sum_{k=1}^{M_n} P\Bigg(\left|\sum_{i=1}^{n}\frac{R_i}{\pi_i(\eta_0)}D_i(\theta_k^*, L\sqrt{b_n})\right| \\
& + \sup_{||\theta|\ \leq 1}\left|\sum_{i=1}^{n}\frac{R_i}{\pi_i(\eta_0)}\left(D_i(\theta, L\sqrt{b_n}) - D_i(\theta_k^*, L\sqrt{b_n})\right)\right| > b_n\omega/2, A_{n1}\cap A_{n2}\Bigg)
\end{aligned}$$

Let $I(\cdot)$ denote the indicator function, we will next show that

$$\sup_{\theta\in\Theta_k}\left|b_n^{-1}\sum_{i=1}^{n}\frac{R_i}{\pi_i(\eta_0)}\left[D_i(\theta, L\sqrt{b_n}) - D_i(\theta_k^*, L\sqrt{b_n})\right]\right| I\left(A_{n1}\cap A_{n2}\right) < \omega/4.$$

Using (7.3), the triangle inequality, condition 6 and the earlier derived bounds for $||\tilde{x}_i||$| and $||\tilde{W}(z_i)||$, we have

$$
\begin{aligned}
& \sup_{\theta\in\Theta_k}\left|b_n^{-1}\sum_{i=1}^{n}\frac{R_i}{\pi_i(\eta_0)}\left(D_i(\theta,L\sqrt{b_n})-D_i(\theta_k^*,L\sqrt{b_n})\right)\right| I\left(A_{n1}\cap A_{n2}\right) \\
= \; & b_n^{-1}\sup_{\theta\in\Theta_k}\left|\sum_{i=1}^{n}\frac{R_i}{\pi_i(\eta_0)}\frac{1}{2}\left[\left|\epsilon_i-\tilde{x}_i'\theta_1 L\sqrt{b_n}-\tilde{W}(z_i)'\theta_2 L\sqrt{b_n}-u_{ni}\right|-\left|\epsilon_i-u_{ni}\right|\right]\right. \\
& -\sum_{i=1}^{n}\frac{R_i}{\pi_i(\eta_0)}\frac{1}{2}E_s\left[\left|\epsilon_i-\tilde{x}_i'\theta_1 L\sqrt{b_n}-\tilde{W}(z_i)'\theta_2 L\sqrt{b_n}-u_{ni}\right|-\left|\epsilon_i-u_{ni}\right|\right] \\
& +\sum_{i=1}^{n}\frac{R_i}{\pi_i(\eta_0)}L\sqrt{b_n}\left(\tilde{x}_i'\theta_1+\tilde{W}(z_i)'\theta_2\right)\psi_\tau(\epsilon_i) \\
& -\sum_{i=1}^{n}\frac{R_i}{\pi_i(\eta_0)}\frac{1}{2}\left[\left|\epsilon_i-\tilde{x}_i'\theta_{k1}^* L\sqrt{b_n}-\tilde{W}(z_i)'\theta_{k2}^* L\sqrt{b_n}-u_{ni}\right|-\left|\epsilon_i-u_{ni}\right|\right] \\
& +\sum_{i=1}^{n}\frac{R_i}{\pi_i(\eta_0)}\frac{1}{2}E_s\left[\left|\epsilon_i-\tilde{x}_i'\theta_{k1}^* L\sqrt{b_n}-\tilde{W}(z_i)'\theta_{k2}^* L\sqrt{b_n}-u_{ni}\right|-\left|\epsilon_i-u_{ni}\right|\right] \\
& \left.-\sum_{i=1}^{n}L\frac{R_i}{\pi_i(\eta_0)}\sqrt{b_n}\left(\tilde{x}_i'\theta_{k1}^*+\tilde{W}(z_i)'\theta_{k2}^*\right)\psi_\tau(\epsilon_i)\right| I\left(A_{n1}\cap A_{n2}\right) \\
\leq \; & 2nLm_0 b_n^{-1/2}\alpha_l^{-1}\max_i\left[||\tilde{x}_i||+||\tilde{W}(z_i)||\right] I\left(A_{n1}\cap A_{n2}\right) \\
\leq \; & 2\alpha_l^{-1}C_2 nLm_0 b_n^{-1/2}\sqrt{b_n/n}=2\alpha_l^{-1}C_2 L\sqrt{n}m_0<\omega/4,
\end{aligned}
$$

by the definition of m_0.

Therefore, to prove $\lim_{n\to\infty}P_{n1}=0$, we only need to verify

$$
\sum_{k=1}^{M_n}P\left(\left|\sum_{i=1}^{n}\frac{R_i}{\pi_i(\eta_0)}D_i(\theta_k^*,L\sqrt{b_n})\right|>b_n\omega/4,A_{n1}\cap A_{n2}\right)\to 0. \tag{7.5}
$$

Applying (7.3), condition 6 and the triangle inequality,

$$\begin{aligned}
&\max_i \left| \frac{R_i}{\pi_i(\eta_0)} D_i(\theta_k^*, L\sqrt{b_n}) \right| I\left(A_{n1} \cap A_{n2}\right) \\
\leq\ & \alpha_l^{-1} \max_i \left| \left| \epsilon_i - \tilde{x}_i'\theta_{k1}^* L\sqrt{b_n} - \tilde{W}(z_i)'\theta_{k2}^* L\sqrt{b_n} - u_{ni} \right| - \left| \epsilon_i - u_{ni} \right| \right| I\left(A_{n1} \cap A_{n2}\right) \\
& + \alpha_l^{-1} \max_i \left| L\sqrt{b_n} \left(\tilde{x}_i'\theta_{k1}^* + \tilde{W}(z_i)'\theta_{k2}^* \right) \psi_\tau(\epsilon_i) \right| I\left(A_{n1} \cap A_{n2}\right) \\
\leq\ & 2\alpha_l^{-1} L\sqrt{b_n} \max_i ||\tilde{s}_i|| I\left(A_{n1} \cap A_{n2}\right) \\
\leq\ & C b_n n^{-1/2},
\end{aligned}$$

for some positive constant C. Define

$$V_i(\theta_k^*, a_n) = Q_i^*(a_n) - Q_i^*(0) + a_n(\tilde{x}_i'\theta_{k1}^* + \tilde{W}(z_i)'\theta_{k2}^*)\psi_\tau(\epsilon_i).$$

Notice that $D_i(\theta_k^*, a_n) = V_i(\theta_k^*, a_n) - E\left[V_i(\theta_k^*, a_n)|x_i,\ z_i\right]$, and that

$$\begin{aligned}
& \sum_{i=1}^n \mathrm{Var}\left(\frac{R_i}{\pi_i(\eta_0)} D_i(\theta_k^*, a_n) I\left(A_{n1} \cap A_{n2}\right) |x_i, z_i \right) \\
\leq\ & \alpha_l^{-2} \sum_{i=1}^n E\left[V_i(\theta_k^*, a_n)^2 I\left(A_{n1} \cap A_{n2}\right) |x_i, z_i \right].
\end{aligned}$$

Using Knight's Identity (Knight 1998) with (7.1),

$$\begin{aligned}
& V_i(\theta_k^*, L\sqrt{b_n}) \\
=\ & L\sqrt{b_n} \left(\tilde{x}_i'\theta_{k1}^* + \tilde{W}(z_i)'\theta_{k2}^* \right) \left[I(\epsilon_i - u_{ni} < 0) - I(\epsilon_i < 0) \right] \\
& + \int_0^{\sqrt{b_n} L\left(\tilde{x}_i'\theta_{k1}^* + \tilde{W}(z_i)'\theta_{k2}^*\right)} \left[I(\epsilon_i - u_{ni} < s) - I(\epsilon_i - u_{ni} < 0) \right] ds \\
\equiv\ & V_{i1} + V_{i2}.
\end{aligned}$$

Using condition 1, we have

$$\begin{aligned}
&\sum_{i=1}^{n} E\left[V_{i1}^2 I\left(A_{n1}\cap A_{n2}\right)|x_i,z_i\right]\\
=\;&\sum_{i=1}^{n} E\left[b_n L^2(\tilde{x}_i'\theta_{k1}^*+\tilde{W}(z_i)'\theta_{k2}^*)^2\left|I(\epsilon_i-u_{ni}<0)-I(\epsilon_i<0)\right| I\left(A_{n1}\cap A_{n2}\right)\Big|x_i,z_i\right]\\
\leq\;& 2L^2 b_n\sum_{i=1}^{n} E\left[\left(\tilde{x}_i'\theta_{k1}^*\right)^2+\left(\tilde{W}(z_i)'\theta_{k2}^*\right)^2)I\left(0\leq|\epsilon_i|\leq|u_{ni}|\right) I\left(A_{n1}\cap A_{n2}\right)\Big|x_i,z_i\right]\\
\leq\;& Cb_n\max_i\left[||\tilde{x}_i||^2+||\tilde{W}(z_i)||^2\right]\sum_{i=1}^{n}\int_{-|u_{ni}|}^{|u_{ni}|} f_i(s)dsI\left(A_{n1}\cap A_{n2}\right)\\
\leq\;& Cb_n^2 J_n^{-r},
\end{aligned}$$

for some positive constant C. Using conditions 1 and 2, we have

$$\begin{aligned}
&\sum_{i=1}^{n} E\left[V_{i2}^2 I\left(A_{n1}\cap A_{n2}\right)\Big|x_i,z_i\right]\\
\leq\;&\max_i\left|\sqrt{b_n}L\left(\tilde{x}_i'\theta_{k1}^*+\tilde{W}(z_i)'\theta_{k2}^*\right)\right|\\
\times\;&\sum_{i=1}^{n}\int_0^{\sqrt{b_n}L\left(\tilde{x}_i'\theta_{k1}^*+\tilde{W}(z_i)'\theta_{k2}^*\right)}\left[F_i(s+u_{ni})-F_i(u_{ni})\right]dsI\left(A_{n1}\cap A_{n2}\right)\\
\leq\;& Cb_n n^{-1/2}\sum_{i=1}^{n}\int_0^{\sqrt{b_n}L\left(\tilde{x}_i'\theta_{k1}^*+\tilde{W}(z_i)'\theta_{k2}^*\right)}(f_i(0)s+f_i'(s^*)s^2)dsI\left(A_{n1}\cap A_{n2}\right)\\
\leq\;& Cb_n^2 n^{-1/2}\left[\theta_{k1}^{*\prime}(\sum_{i=1}^{n}\tilde{x}_i\tilde{x}_i')\theta_{k1}^*+\theta_{k2}^{*\prime}(\sum_{i=1}^{n}\tilde{W}(z_i)\tilde{W}(z_i))\theta_{k2}^*\right](1+o(1))\\
\leq\;& Cb_n^2 n^{-1/2}\left[||\theta_{k1}^*||^2\lambda_{\max}(n^{-1}X^{*\prime}X^*)\right.\\
&\left.+||\theta_{k2}^*||^2||W_B^{-1}||^2\lambda_{\max}(W'W)\right](1+o(1))\\
\leq\;& Cb_n^2 n^{-1/2}(1+o(1)),
\end{aligned}$$

for some positive constant C, where the second to last inequality applies condition 2 and the result of Zhou et al. (1998) on the properties of basis functions. Therefore

$$\sum_{i=1}^{n} \mathrm{Var}\Big(D_i(\theta) I\left(A_{n1} \cap A_{n2}\right) \Big| x_i, z_i\Big) \leq C b_n^2 n^{-1/2},$$

for some positive constant C and all n sufficiently large. By Bernstein's inequality, for all n sufficiently large,

$$\begin{aligned}
& \sum_{k=1}^{M_n} P\left(\left|\sum_{i=1}^{n} D_i(\theta_k^*, L\sqrt{b_n/n})\right| > b_n\omega/2, A_{n1} \cap A_{n2} \Big| x_i, z_i\right) \\
\leq\ & 2\sum_{k=1}^{M_n} \exp\left(\frac{-b_n^2\omega^2/4}{Cb_n^2 n^{-1/2} + C\omega b_n^2 n^{-1/2}}\right) \\
\leq\ & 2\sum_{k=1}^{M_n} \exp\left(-C\sqrt{n}\right) \\
=\ & 2M_n \exp\left(-C\sqrt{n}\right) \\
\leq\ & C\left(\frac{C\sqrt{n}}{\omega}\right)^{b_n+1} \exp\left(-C\sqrt{n}\right) \\
=\ & C\exp\left((b_n+1)\log(C\sqrt{n}/\omega) - C\sqrt{n}\right) \\
\leq\ & C\exp\left(C(b_n+1)\log(n) - C\sqrt{n}\right),
\end{aligned}$$

which converges to zero as $n \to \infty$. Note that the upper bound does not depend on $\{x_i, z_i\}$. This implies $\lim_{n\to\infty} P_{n1} = 0$.

Define $D(\theta, a_n) = (D_1(\theta, a_n), ..., D_n(\theta, a_n))' \in \mathbb{R}^n$, $R = (R_1, ..., R_n) \in \mathcal{R}^n$ and $\pi(\eta)^{-1} = \mathrm{diag}\left(\pi_1(\eta)^{-1}, ..., \pi_n(\eta)^{-1}\right) \in \mathcal{R}^{n\times n}$. Notice

$$P_{n2} = P\left(\sup_{\substack{||\theta|| \leq 1 \\ ||\eta-\eta_0|| \leq Cn^{-1/2}}} \left|b_n^{-1} D'(\pi(\eta)^{-1} - \pi(\eta_0)^{-1})R\right| > \omega/2, A_{n1} \cap A_{n2}\right)$$

Using the same methods that showed $\lim_{n\to\infty} P_{n1} = 0$ we have $\sup_{||\theta|| \leq 1} |b_n^{-1} D'R| = o_p(1)$. Using conditions 6 and 7

$$\begin{aligned} & \sup_{||\eta-\eta_0||\leq Cn^{-1/2}} \max_i \left(\frac{1}{\pi_i \eta} - \frac{1}{\pi_i(\eta_0)} \right) \\ = & \sup_{||\eta-\eta_0||\leq Cn^{-1/2}} \max_i (\eta - \eta_0)' \left(\frac{\partial \pi_i(\eta)}{\partial \eta} \right)_{\eta=\eta_0} \frac{1}{\pi_i(\eta_0)^2} (1 + o_p(1)) = o_p(1). \end{aligned}$$

Proof is complete because $\lim_{n\to\infty} P_{n1} = 0$ and $\lim_{n\to\infty} P_{n2} = 0$. □

Lemma 5

If the conditions of Theorem 4.3 hold, then for any $\omega > 0$ there exists an $L > 0$ such that

$$P\left(\inf_{||\theta||=L} b_n^{-1} \sum_{i=1}^n \frac{R_i}{\pi_i(\hat{\eta})} (Q_i(\sqrt{b_n}) - Q_i(0)) > 0 \right) \geq 1 - \omega. \qquad \square$$

Proof: Note that

$$\begin{aligned} b_n^{-1} \sum_{i=1}^n \frac{R_i}{\pi_i(\hat{\eta})} (Q_i(\sqrt{b_n}) - Q_i(0)) &= b_n^{-1} \sum_{i=1}^n \frac{R_i}{\pi_i(\hat{\eta})} D_i(\theta, \sqrt{b_n}) \\ &+ b_n^{-1} \sum_{i=1}^n \frac{R_i}{\pi_i(\hat{\eta})} E_s[Q_i(\sqrt{b_n}) - Q_i(0)] \\ &- b_n^{-1/2} \sum_{i=1}^n \frac{R_i}{\pi_i(\hat{\eta})} (\tilde{x}_i' \theta_1 + \tilde{W}(z_i)' \theta_2) \psi_\tau(\epsilon_i) \\ &= D_{n1} + D_{n2} + D_{n3}, \end{aligned}$$

where the definition of D_{ni}, $i = 1, 2, 3$, is clear from the context. Lemma 4 and condition 7 provide that $D_{n1} = o_p(1)$. We next evaluate D_{n3}. First

$$\begin{aligned} D_{n3} &= b_n^{-1/2} \sum_{i=1}^{n} \frac{R_i}{\pi_i(\eta_0)} (\tilde{x}_i'\theta_1 + \tilde{W}(z_i)'\theta_2)\psi_\tau(\epsilon_i) \\ &+ b_n^{-1/2} \sum_{i=1}^{n} R_i \left(\frac{1}{\pi_i(\hat{\eta})} - \frac{1}{\pi_i(\eta_0)} \right) (\tilde{x}_i'\theta_1 + \tilde{W}(z_i)'\theta_2)\psi_\tau(\epsilon_i) \\ &= D_{n31} + D_{n32}. \end{aligned}$$

Note that $E(D_{n31}) = 0$ and by 3 and conditions 2 and 7

$$\begin{aligned} E(D_{n31}^2) &\leq C b_n^{-1} \mathrm{E}\left[n^{-1}\theta_1' X^{*\prime} X^* \theta_1 + ||W_B^{-1}||^2 \theta_2 W' W \theta_2 \right] \\ &= O\left(b_n^{-1} ||\theta||^2 \right). \end{aligned}$$

Let $\partial\pi(\gamma_0) = \left(\left(\frac{\partial \pi_1(\eta)}{\partial \eta} \right)'_{\eta=\eta_0}, ..., \left(\frac{\partial \pi_n(\eta)}{\partial \eta} \right)'_{\eta=\eta_0} \right)'$. Using condition 7, the rate for W_B^{-1} and Taylor expansion

$$\begin{aligned} D_{n32} &= b_n^{-1/2} (\hat{\eta} - \eta_0)' \sum_{i=1}^{n} \left(\frac{\partial \pi_i(\eta)}{\partial \eta} \right)_{\eta=\eta_0} \frac{R_i}{\pi_i(\eta_0)^2} (\tilde{x}_i'\theta_1 + \tilde{W}(z_i)'\theta_2)\psi_\tau(\epsilon_i) \\ &= b_n^{-1/2} \sqrt{n} (\hat{\eta} - \eta_0)' n^{-1} \sum_{i=1}^{n} \left(\frac{\partial \pi_i(\eta)}{\partial \eta} \right)_{\eta=\eta_0} \frac{R_i}{\pi_i(\eta_0)^2} \delta_i' \theta_1 \psi_\tau(\epsilon_i)(1 + o(1)) \\ &+ b_n^{-1/2} (\hat{\eta} - \eta_0)' \sum_{i=1}^{n} \left(\frac{\partial \pi_i(\eta)}{\partial \eta} \right)_{\eta=\eta_0} \frac{R_i}{\pi_i(\eta_0)^2} \psi_\tau(\epsilon_i) W(z_i)' W_B^{-1} \theta_2 \\ &= O\left(b_n^{-1/2} ||\theta|| \right). \end{aligned}$$

Therefore $D_{n3} = O_p\left(b_n^{-1/2}||\theta||\right)$. Before analyzing D_{n2} we present some results to assist in understanding its asymptotic behavior. Let

$$\begin{aligned}
& \frac{1}{2}\sum_{i=1}^{n}\frac{R_i}{\pi_i(\hat{\eta})}f_i(0)\left(\tilde{x}_i'\theta_1+\tilde{W}(z_i)'\theta_2\right)^2 \\
= & \frac{1}{2}\sum_{i=1}^{n}\frac{R_i}{\pi_i(\eta_0)}f_i(0)\left(\tilde{x}_i'\theta_1+\tilde{W}(z_i)'\theta_2\right)^2 \\
+ & \frac{1}{2}\sum_{i=1}^{n}R_i\left(\frac{1}{\pi_i(\hat{\eta})}-\frac{1}{\pi_i(\eta_0)}\right)f_i(0)\left(\tilde{x}_i'\theta_1+\tilde{W}(z_i)'\theta_2\right)^2 \\
= & E_{n1}+E_{n2},
\end{aligned}$$

where definition of E_{n1} and E_{n2} is immediate from the context. We further separate E_{n1} with

$$\begin{aligned}
E_{n1} &= \frac{1}{2}\sum_{i=1}^{n}\frac{R_i}{\pi_i(\eta_0)}f_i(0)(\tilde{x}_i'\theta_1)^2 \\
&+ \frac{1}{2}\sum_{i=1}^{n}\frac{R_i}{\pi_i(\eta_0)}f_i(0)(\tilde{W}(z_i)'\theta_2)^2 \\
&+ \sum_{i=1}^{n}\frac{R_i}{\pi_i(\eta_0)}f_i(0)(\tilde{x}_i'\theta_1)(\tilde{W}(z_i)'\theta_2) \\
&= E_{n11}+E_{n12}+E_{n13}.
\end{aligned}$$

By definition of $\tilde{x}_i$, $\tilde{W}(z_i)$ and condition 6

$$\begin{aligned}
E[E_{n13}] &= E\left[\frac{1}{2}\sum_{i=1}^{n}f_i(0)\tilde{x}_i'\theta_1\tilde{W}(z_i)'\theta_2\right]=0, \\
\text{Var}\left(E_{n13}\right) &\leq \alpha_l^{-2}\text{Var}\left(\frac{1}{2}\sum_{i=1}^{n}f_i(0)\tilde{x}_i'\theta_1\tilde{W}(z_i)'\theta_2\right)=0.
\end{aligned}$$

Using condition 7 to analyze E_{n2} we have

$$\begin{aligned}
E_{n2} &= \frac{1}{2}(\hat{\eta}-\eta_0)'\sum_{i=1}^{n}\left(\frac{\partial\pi_i(\gamma)}{\partial\gamma}\right)_{\eta=\eta_0}\frac{R_i}{\pi_i(\eta_0)^2}f_i(0)\left(\tilde{x}_i'\theta_1+\tilde{W}(z_i)'\theta_2\right)^2(1+o_p(1)) \\
&\leq (\hat{\eta}-\eta_0)'\sum_{i=1}^{n}\left(\frac{\partial\pi_i(\gamma)}{\partial\gamma}\right)_{\eta=\eta_0}\frac{R_i}{\pi_i(\eta_0)^2}f_i(0)\left(\tilde{x}_i'\theta_1\right)^2(1+o_p(1)) \\
&+ (\hat{\eta}-\eta_0)'\sum_{i=1}^{n}\left(\frac{\partial\pi_i(\gamma)}{\partial\gamma}\right)_{\eta=\eta_0}\frac{R_i}{\pi_i(\eta_0)^2}f_i(0)\left(\tilde{W}(z_i)'\theta_2\right)^2(1+o(1)) \\
&= O_p\left(n^{-1/2}\right).
\end{aligned}$$

Applying Knight's identity (Knight, 1998), using condition 1 and noting

$$\frac{1}{2}\sum_{i=1}^{n}\frac{R_i}{\pi_i(\hat{\eta})}f_i(0)\left(\tilde{x}_i'\theta_1+\tilde{W}(z_i)'\theta_2\right)^2=E_{n1}+E_{n2}+o_p(1),$$

we have

$$\begin{aligned}
D_{n2} &= b_n^{-1}\sum_{i=1}^{n}\frac{R_i}{\pi_i(\hat{\eta})}E\left[\int_{-u_{ni}}^{-\sqrt{b_n}\left(\tilde{x}_i'\theta_1+\tilde{W}(z_i)'\theta_2\right)-u_{ni}}\psi_\tau(\epsilon_i+s)ds\middle|x_i,z_i\right] \\
&= b_n^{-1}\sum_{i=1}^{n}\frac{R_i}{\pi_i(\hat{\eta})}\int_{u_{ni}}^{\sqrt{b_n}\left(\tilde{x}_i'\theta_1+\tilde{W}(z_i)'\theta_2\right)+u_{ni}}f_i(0)s\,ds(1+o(1)) \\
&= \theta_1'\left(n^{-1}\frac{1}{2}\sum_{i=1}^{n}\frac{R_i}{\pi_i(\eta_0)}f_i(0)x_i^*x_i^{*\prime}\right)\theta_1(1+o(1)) \\
&+ \theta_2'\left(\frac{1}{2}\sum_{i=1}^{n}\frac{R_i}{\pi_i(\eta_0)}f_i(0)\tilde{W}(z_i)\tilde{W}(z_i)'\right)\theta_2(1+o(1)) \\
&+ b_n^{-1/2}\sum_{i=1}^{n}\frac{R_i}{\pi_i(\hat{\eta})}f_i(0)u_{ni}\left(\tilde{x}_i'\theta_1+\tilde{W}(z_i)'\theta_2\right)+o(1).
\end{aligned}$$

By Lemma 3 and condition 5, $n^{-1}\theta_1'X^{*\prime}B_nX^*\theta_1 = \theta_1'\Sigma_1\theta_1+o_p(1)$. Therefore by adapting results from Zhou et al. (1998) to handle the B-spline basis terms there

exists a finite positive constant c, such that with probability approaching one

$$n^{-1}\theta_1' \sum_{i=1}^{n} \frac{R_i}{\pi_i(\eta_0)} f_i(0) x_i^* x_i^{*\prime} \theta_1 + \theta_2 W_B^{-1} \sum_{i=1}^{n} \frac{R_i}{\pi_i(\eta_0)} f_i(0) W(z_i) W(z_i)' W_B^{-1} \theta_2 \geq c||\theta||^2.$$

Define $U_n = (u_{n1}, ..., u_{nn})'$. Then, by Schumaker (1981), $||U_n|| = O_p\left(\sqrt{n} J_n^{-r}\right) = o_p(1)$. Define $\hat{R}_n = \text{diag}\left(R_1 \pi_1(\hat{\eta})^{-1}, ..., R_n \pi_n(\hat{\eta})^{-1}\right)$. Using results shown in Lemma 4 $||\hat{R}|| = O_p(1)$. Thus, for the linear terms,

$$\begin{aligned} b_n^{-1/2} \sum_{i=1}^{n} \frac{R_i}{\pi_i(\hat{\eta})} f_i(0) u_{ni} \tilde{x}_i' \theta_1 &= b_n^{-1/2} n^{-1/2} \theta_1' X^{*\prime} B_n \hat{R}_n U_n = o_p(||\theta||), \\ b_n^{-1/2} \sum_{i=1}^{n} \frac{R_i}{\pi_i(\hat{\eta})} f_i(0) u_{ni} \tilde{W}(z_i)' \theta_2 &= b_n^{-1/2} \theta_2' W_B^{-1} W' B_n \hat{R}_n U_n = o_p(||\theta||). \end{aligned}$$

For L sufficiently large, the always positive quadratic term asymptotically dominates. This proves the lemma. □

7.1.5 Lemmas for proof of Theorem 4.4

In some of the lemmas we use the following definition for ease of notation

$$Q_i^*(\theta_1, \tilde{\theta}_1, \theta_2) = \rho_\tau(\epsilon_i - \tilde{x}_i'\theta_1 - \tilde{W}(z_i)'\theta_2 - u_{ni}) - \rho_\tau(\epsilon_i - \tilde{x}_i'\tilde{\theta}_1 - \tilde{W}(z_i)'\theta_2 - u_{ni}).$$

Lemma 6

If the conditions of Theorem 4.4 hold, then

$$\sup_{\substack{||\theta_1 - \tilde{\theta}_1|| \leq M \\ ||\theta_2|| \leq C\sqrt{b_n}}} \left| \sum_{i=1}^{n} \frac{R_i}{\pi_i(\hat{\eta})} E_s\left[Q_i^*(\theta_1, \tilde{\theta}_1, \theta_2)\right] - \frac{1}{2}\left[\theta_1' \Sigma_1 \theta_1 - \tilde{\theta}_1' \Sigma_1 \tilde{\theta}_1\right](1 + o(1)) \right| = o_p(1). \quad \square$$

Proof: Applying Knight's formula (Knight, 1998) and methods used in the proof of Theorem 4.1 and Lemma 5 to handle the weights, we have

$$\begin{aligned}
& \sum_{i=1}^{n} \frac{R_i}{\pi_i(\hat{\eta})} E_s \left[Q_i^*(\theta_1, \tilde{\theta}_1, \theta_2) \right] \\
= & \sum_{i=1}^{n} \frac{R_i}{\pi_i(\hat{\eta})} \int_{\tilde{x}_i'\tilde{\theta}_1 + \tilde{W}(z_i)'\theta_2 + u_{ni}}^{\tilde{x}_i'\theta_1 + \tilde{W}(z_i)'\theta_2 + u_{ni}} (F_i(s) - F_i(0)) ds \\
= & \frac{1}{2} \sum_{i=1}^{n} \frac{R_i}{\pi_i(\hat{\eta})} f_i(0)(1 + o(1)) \left[\left(\tilde{x}_i'\theta_1 + \tilde{W}(z_i)'\theta_2 + u_{ni} \right)^2 - \left(\tilde{x}_i'\tilde{\theta}_1 + \tilde{W}(z_i)'\theta_2 + u_{ni} \right)^2 \right] \\
= & \frac{1}{2} \sum_{i=1}^{n} \frac{R_i}{\pi_i(\hat{\eta})} f_i(0) \left[(\tilde{x}_i'\theta_1)^2 - \left(\tilde{x}_i'\tilde{\theta}_1 \right)^2 + 2 \left(\theta_2' \tilde{W}(z_i) + u_{ni} \right) \left(\tilde{x}_i'\theta_1 - \tilde{x}_i'\tilde{\theta}_1 \right) \right] (1 + o(1)) \\
= & \frac{1}{2} \left[\theta_1' \Sigma_1 \theta_1 - \tilde{\theta}_1' \Sigma_1 \tilde{\theta}_1 \right] (1 + o(1)) + \frac{1}{2} n^{-1/2} (\theta_1 - \tilde{\theta}_1)' X^{*'} B_n \hat{R} U_n (1 + o(1)).
\end{aligned}$$

The proof is complete by noting that $\sup_{||\theta_1 - \tilde{\theta}_1|| \leq M,\, ||\theta_2|| \leq C\sqrt{b_n}} |n^{-1/2}(\theta_1 - \tilde{\theta}_1)' X^{*'} B_n \hat{R} U_n (1 + o(1))| = o_p(1)$. □

Lemma 7

If the conditions of Theorem 4.4 hold, then for any given positive constants M and C,

$$\sup_{\substack{||\theta_1 - \tilde{\theta}_1|| \leq M \\ ||\theta_2|| \leq C\sqrt{b_n}}} \left| \sum_{i=1}^{n} \frac{R_i}{\pi_i(\hat{\eta})} \left[Q_i^*(\theta_1, \tilde{\theta}_1, \theta_2) - E_s \left[Q_i^*(\theta_1, \tilde{\theta}_1, \theta_2) \right] + \tilde{x}_i' \left(\theta_1 - \tilde{\theta}_1 \right) \psi_\tau(\epsilon_i) \right] \right| = o_p(1).$$

□

Proof: Define $A_i(\theta_1, \tilde{\theta}_1, \theta_2) = Q_i^*(\theta_1, \tilde{\theta}_1, \theta_2) - E_s \left[Q_i^*(\theta_1, \tilde{\theta}_1, \theta_2) \right] + \tilde{x}_i' \left(\theta_1 - \tilde{\theta}_1 \right) \psi_\tau(\epsilon_i)$. Let A_{n1} and A_{n2} be defined as in Lemma 4. We separate the problem into solving

two probability statements. Note that for any given $\omega > 0$

$$\begin{aligned}
& P\left(\sup_{||\theta_1-\tilde{\theta}_1||\leq M,\ ||\theta_2||\leq C\sqrt{b_n}}\left|\sum_{i=1}^{n}\frac{R_i}{\pi_i(\eta_0)}A_i(\theta_1,\tilde{\theta}_1,\theta_2)\right| > \omega,\ A_{n1}\cap A_{n2}\right) \\
\leq\ & P\left(\sup_{||\theta_1-\tilde{\theta}_1||\leq M,\ ||\theta_2||\leq C\sqrt{b_n}}\left|\sum_{i=1}^{n}\frac{R_i}{\pi_i(\eta_0)}A_i(\theta_1,\tilde{\theta}_1,\theta_2)\right| > \omega/2,\ A_{n1}\cap A_{n2}\right) \\
+\ & P\left(\sup_{||\theta_1-\tilde{\theta}_1||\leq M,\ ||\theta_2||\leq C\sqrt{b_n}}\left|\sum_{i=1}^{n}R_i\left(\frac{1}{\pi_i(\hat{\eta})}-\frac{1}{\pi_i(\eta_0)}\right)A_i(\theta_1,\tilde{\theta}_1,\theta_2)\right|\right. \\
& \left. > \omega/2,\ A_{n1}\cap A_{n2}\right) \\
=\ & P^*_{n1}+P^*_{n2}.
\end{aligned}$$

Where the definition of P^*_{n1} and P^*_{n2} follows directly from the context. Then lemma is proved if we show $P_{n1}\to 0$ and $P_{n2}\to 0$. We first work with P_{n1}. We note that

$$\begin{aligned}
A_i(\theta_1,\tilde{\theta}_1,\theta_2) =\ & \frac{1}{2}\left[|\epsilon_i-\tilde{x}_i'\theta_1-\tilde{W}(z_i)'\theta_2-u_{ni}|-|\epsilon_i-\tilde{x}_i'\tilde{\theta}_1-\tilde{W}(z_i)'\theta_2-u_{ni}|\right] \\
& +(\tau-1/2)(\tilde{x}_i'(\tilde{\theta}_1-\theta_1)) \\
& -\frac{1}{2}E_s\left[|\epsilon_i-\tilde{x}_i'\theta_1-\tilde{W}(z_i)'\theta_2-u_{ni}|-|\epsilon_i-\tilde{x}_i'\tilde{\theta}_1-\tilde{W}(z_i)'\theta_2-u_{ni}|\right] \\
& -E_s\left[(\tau-1/2)(\tilde{x}_i'(\tilde{\theta}_1-\theta_1))\right] \\
& +\tilde{x}_i'(\theta_1-\tilde{\theta}_1)\psi_\tau(\epsilon_i).
\end{aligned}$$

Similarly as in the proof of Lemma 4, let $\Theta_1=\left\{\theta_1:\ ||\theta-\tilde{\theta}_1||\leq M,\ \theta_1\in\mathbb{R}^{p+1}\right\}$ and $\Theta_2=\left\{\theta_2:\ ||\theta_2||\leq C\sqrt{b_n},\ \theta_2\in\mathbb{R}^{dJ_n}\right\}$. We can partition Θ_1 (similarly Θ_2), into disjoint regions $\Theta_{11},...,\Theta_{1K_n}$ $(\Theta_{21},...,\Theta_{2L_n})$ such that the diameter of each region does not exceed $m_0^*=\frac{C\omega}{4\sqrt{nb_n}}$. These partitions can be constructed such that $K_n\leq C\left(\frac{C\sqrt{nb_n}}{4\omega}\right)^{p+1}$ and $L_n\leq C\left(\frac{C\sqrt{nb_n}}{4\omega}\right)^{dJ_n}$. Let $\theta^*_{11},...,\theta^*_{1K_n}$ be arbitrary points in $\Theta_{11},...,\Theta_{1K_n}$, respectively; similarly, let $\theta^*_{21},...,\theta^*_{2L_n}$ be arbitrary points in $\Theta_{21},...,\Theta_{2L_n}$, respectively.

Then

$$P^*_{n1} \leq \sum_{l=1}^{L_n}\sum_{k=1}^{K_n} P\Bigg(\sup_{\theta_1\in\Theta_{1k}}\sup_{\theta_2\in\Theta_{2l}}\Big|\sum_{i=1}^{n}\frac{R_i}{\pi_i(\eta_0)}A_i(\theta_{1k},\tilde{\theta}_1,\theta_{2l})+A_i(\theta_1,\tilde{\theta}_1,\theta_2) - A_i(\theta_{1k},\tilde{\theta}_1,\theta_{2l})\Big|, > \omega/2\, A_{n1}\cap A_{n2}\Bigg).$$

Note that

$$\begin{aligned}
& Q^*_i(\theta_1,\tilde{\theta}_1,\theta_2) - Q^*_i(\theta_{1k},\tilde{\theta}_1,\theta_{2l}) \\
= & \frac{1}{2}\left[\left|\epsilon_i - \tilde{x}_i'\theta_1 - \tilde{W}(z_i)'\theta_2 - u_{ni}\right| - \left|\epsilon_i - \tilde{x}_i'\tilde{\theta}_1 - \tilde{W}(z_i)'\theta_2 - u_{ni}\right|\right] \\
- & \frac{1}{2}\left[\left|\epsilon_i - \tilde{x}_i'\theta_{1k} - \tilde{W}(z_i)'\theta_{2l} - u_{ni}\right| - \left|\epsilon_i - \tilde{x}_i'\tilde{\theta}_1 - \tilde{W}(z_i)'\theta_{2l} - u_{ni}\right|\right] \\
+ & (\tau-1/2)(\tilde{x}_i'(\tilde{\theta}_1-\theta_1)) - (\tau-1/2)(\tilde{x}_i'(\tilde{\theta}-\theta_{1k})) \\
= & \frac{1}{2}\left[\left|\epsilon_i - \tilde{x}_i'\theta_1 - \tilde{W}(z_i)'\theta_2 - u_{ni}\right| - \left|\epsilon_i - \tilde{x}_i'\theta_{1k} - \tilde{W}(z_i)'\theta_{2l} - u_{ni}\right|\right] \\
- & \frac{1}{2}\left[\left|\epsilon_i - \tilde{x}_i'\tilde{\theta}_1 - \tilde{W}(z_i)'\theta_2 - u_{ni}\right| - \left|\epsilon_i - \tilde{x}_i'\tilde{\theta}_{1k} - \tilde{W}(z_i)'\theta_{2l} - u_{ni}\right|\right] \\
+ & (\tau-1/2)(\tilde{x}_i'(\theta_{1k}-\theta_1)) \\
\leq & 2\max_i||\tilde{s}_i||\sup_{\theta_1\in\Theta_{1k}}\sup_{\theta_2\in\Theta_{2l}}[||\theta_1-\theta_{1k}|| + ||\theta_2-\theta_{2l}||]
\end{aligned}$$

Using the above inequality, the definition of $A_i(\theta_1,\tilde{\theta}_1,\theta_2)$, m^*_0 and conditions 2 and 6

$$\begin{aligned}
& \sup_{\theta_1\in\Theta_{1k}}\sup_{\theta_2\in\Theta_{2l}}\sum_{i=1}^{n}\frac{R_i}{\pi_i(\eta_0)}\left|A_i(\theta_1,\tilde{\theta}_1,\theta_2) - A_i(\bar{\theta}_{l_1},\tilde{\theta}_1,\bar{\theta}_{l_2})\right| I(A_{n1}\cap A_{n2}) \\
\leq & 5n\max_i||\tilde{s}_i||\sup_{\theta_1\in\Theta_{1k}}\sup_{\theta_2\in\Theta_{2l}}[||\theta_1-\theta_{1k}|| + ||\theta_2-\theta_{2l}||]\, I(A_{n1}\cap A_{n2}) \\
\leq & Cm^*_o\sqrt{nb_n} \leq \omega/4.
\end{aligned}$$

Then $P^*_{n1} \to 0$ if for any $\omega > 0$

$$\sum_{l=1}^{L_n}\sum_{k=1}^{K_n} P\left(\left|\sum_{i=1}^{n}\frac{R_i}{\pi_i(\eta_0)}A_i(\theta_{1k},\tilde{\theta}_1,\theta_{2l})\right| > \omega/4\right) \to 0.$$

Bernstein's inequality will be used and the variance and maximum of $A_i(\theta_{1k},\tilde{\theta}_1,\theta_{2l})$ is needed. Assuming $\tilde{s}_{(n)} < C\sqrt{b_n/n}$ and noting this depends on $\max_i||\tilde{x}_i|| < n^{-1/2}$. Also noting that this lemma requires that there exists a positive constant C such that $||\theta_1 - \tilde{\theta}_1|| < C$ then the maximum has the following upper bound

$$\begin{aligned}
& \max_i \left|\frac{R_i}{\pi_i(\eta_0)}A_i(\theta_{1k},\tilde{\theta}_1,\theta_{2l})\right| \\
= \; & \max_i\left[\frac{R_i}{\pi_i(\eta_0)}\frac{1}{2}\left[|\epsilon_i - \tilde{x}_i'\theta_{1k} - \tilde{W}(z_i)'\theta_{2l} - u_{ni}| - |\epsilon_i - \tilde{x}_i'\tilde{\theta}_1 - \tilde{W}(z_i)'\theta_{2l} - u_{ni}|\right]\right. \\
- \; & \frac{1}{2}E_s\left[|\epsilon_i - \tilde{x}_i'\theta_{1k} - \tilde{W}(z_i)'\theta_{2l} - u_{ni}| - |\epsilon_i - \tilde{x}_i'\tilde{\theta}_1 - \tilde{W}(z_i)'\theta_{2l} - u_{ni}|\right] \\
+ \; & (\tau - 1/2)\left(\tilde{x}_i'(\tilde{\theta}_1 - \theta_{1k})\right) - E_s\left[(\tau - 1/2)\left(\tilde{x}_i'(\tilde{\theta}_1 - \theta_{1k})\right)\right] \\
+ \; & \left.\tilde{x}_i'(\theta_{1k} - \tilde{\theta}_1)\psi_\tau(\epsilon_i)\right] \\
\leq \; & 3\alpha_l^{-1}\max_i||\tilde{x}_i||||\theta_1 - \tilde{\theta}_1|| \\
\leq \; & Cn^{-1/2}.
\end{aligned}$$

Using Knight's identity (Knight, 1998)

$$\begin{aligned}
& \rho_\tau(\epsilon_i - \tilde{x}_i'\theta_{1k} - \tilde{W}(z_i)'\theta_{2l} - u_{ni}) - \rho_\tau(\epsilon_i - \tilde{x}_i'\tilde{\theta}_1 - \tilde{W}(z_i)'\theta_{2l} - u_{ni}) \\
& +(\theta_{1k} - \tilde{\theta}_1)'\tilde{x}_i\psi_\tau(\epsilon_i) \\
= \; & (\theta_{1k} - \tilde{\theta}_1)'\tilde{x}_i\left(I(\epsilon_i < \tilde{x}_i'\tilde{\theta}_1 + \tilde{W}(z_i)'\theta_{2l} + u_{ni}) - I(\epsilon_i < 0)\right) \\
+ \; & \int_0^{\tilde{x}_i'(\theta_{1k}-\tilde{\theta}_1)} I(\epsilon_i \leq s + \tilde{x}_i'\tilde{\theta}_1 + \tilde{W}(z_i)'\theta_{2l} + u_{ni}) - I(\epsilon_i \leq \tilde{x}_i'\tilde{\theta}_1 + \tilde{W}(z_i)'\theta_{2l} + u_{ni})ds \\
= \; & D_{i1} + D_{i2}
\end{aligned}$$

To get an upper bound for $\sum_{i=1}^{n} \text{Var}(\frac{R_i}{\pi_i(\eta_0)} A_i(\theta_{1k}, \tilde{\theta}_1, \theta_{2l}))$ we analyze $\sum_{i=1}^{n} E\left[D_{i1}^2\right]$ and $\sum_{i=1}^{n} E\left[D_{i2}^2\right]$. Using rate of convergence of $\tilde{\theta}_1$, conditions 1 and 5, the definitions of θ_{1k} and θ_{2l}, the rate of $\max_i |u_{ni}|$ and $\max_i ||\tilde{s}_i|| < \sqrt{b_n/n}$. Then

$$
\begin{aligned}
\sum_{i=1}^{n} E\left[D_{i1}^2\right] &= \sum_{i=1}^{n} E\left[\left(\tilde{x}_i'\left(\theta_{1k} - \tilde{\theta}_1\right)\right)^2 \left| I(\epsilon_i < \tilde{x}_i'\tilde{\theta}_1 + \tilde{W}(z_i)'\theta_{2l} + u_{ni}) - I(\epsilon_i < 0)\right|\right] \\
&= \sum_{i=1}^{n} E\left[\left(\tilde{x}_i'\left(\theta_{1k} - \tilde{\theta}_1\right)\right)^2 I\left(0 \leq |\epsilon_i| \leq \left|\tilde{x}_i'\tilde{\theta}_1 + \tilde{W}(z_i)'\theta_{2l} + u_{ni}\right|\right)\right] \\
&= \sum_{i=1}^{n} E\left[\left(\tilde{x}_i'\left(\theta_{1k} - \tilde{\theta}_1\right)\right)^2 \int_{-|\tilde{x}_i'\tilde{\theta}_1 + \tilde{W}(z_i)'\theta_{2l} + u_{ni}|}^{|\tilde{x}_i'\tilde{\theta}_1 + \tilde{W}(z_i)'\theta_{2l} + u_{ni}|} f_i(s)ds\right] \\
&\leq C\sum_{i=1}^{n} E\left[\left(\tilde{x}_i'\left(\theta_{1k} - \tilde{\theta}_1\right)\right)^2 |\tilde{x}_i'\tilde{\theta}_1 + \tilde{W}(z_i)'\theta_{2l} + u_{ni}|\right] \\
&\leq C\max_i \left|\tilde{x}_i'\tilde{\theta}_1 + \tilde{W}(z_i)'\theta_{2l} + u_{ni}\right| \sum_{i=1}^{n} E\left[\left(\tilde{x}_i'\left(\theta_{1k} - \tilde{\theta}_1\right)\right)^2\right] \\
&\leq C\left(\sqrt{b_n}n^{-1/2} + J_n^{-r}\right) E\left[(\theta_{1k} - \tilde{\theta}_1)'\frac{1}{n}\sum_{i=1}^{n} x_i^* x_i^{*\prime}(\theta_{1k} - \tilde{\theta}_1)\right] \\
&\leq C\sqrt{b_n}n^{-1/2}.
\end{aligned}
$$

Using similar techniques for D_{i2}

$$\begin{aligned}
\sum_{i=1}^{n} E\left[D_{i2}^2\right] &\leq \max_i \left|\tilde{x}_i'(\theta_{1k}-\tilde{\theta}_1)\right| \\
&\times \sum_{i=1}^{n} E\left[\int_0^{\tilde{x}_i'(\theta_{1k}-\tilde{\theta}_1)} \left[F_i(s+\tilde{x}_i'\tilde{\theta}_1+\tilde{W}(z_i)'\theta_{2l}+u_{ni})\right.\right. \\
&\quad \left.\left. -F_i(\tilde{x}_i'\tilde{\theta}_1+\tilde{W}(z_i)'\theta_{2l}+u_{ni})ds\right]\right] \\
&\leq Cn^{-1/2}\sum_{i=1}^{n} E\left[\int_0^{\tilde{x}_i'(\theta_{1k}-\tilde{\theta}_1)} sf_i\left(\tilde{x}_i'\tilde{\theta}_1+\tilde{W}(z_i)'\theta_{2l}+u_{ni}\right)ds\right]+o(1) \\
&\leq Cn^{-1/2}\left[\theta_1'\frac{1}{n}\sum_{i=1}^{n} E\left[x_i^* x_i^{*\prime}\right]\theta_1\right] \\
&\leq Cn^{-1/2}.
\end{aligned}$$

Therefore by condition 6

$$\sum_{i=1}^{n} \mathrm{Var}\left(\frac{R_i}{\pi_i(\eta_0)}A_i(\theta_{1k},\tilde{\theta}_1,\theta_{2l})\right) I\left(A_{n1}\cap A_{n2}\right) \leq C\sqrt{\frac{b_n}{n}}.$$

Using Bernstein's inequality and conditions 4

$$\begin{aligned}
& \sum_{l=1}^{L_n}\sum_{k=1}^{K_n} P\left(\left|\sum_{i=1}^{n}\frac{R_i}{\pi_i(\eta_0)}A_i(\theta_{1k},\tilde{\theta}_1,\theta_{2l})\right| > \omega/2 \middle| \tilde{s}_{(n)} < C\sqrt{b_n/n}\right) \\
\leq & \sum_{l=1}^{L_n}\sum_{k=1}^{K_n}\exp\left(\frac{-\omega^2/4}{C\sqrt{b_n}n^{-1/2}+\omega C n^{-1/2}}\right) \\
\leq & \sum_{l=1}^{L_n}\sum_{k=1}^{K_n}\exp\left(-C\sqrt{n}b_n^{-1/2}\right) \\
= & L_n K_n \exp\left(-C\sqrt{n}b_n^{-1/2}\right) \\
\leq & C\left(C\sqrt{nb_n}\right)^{p+1}\left(C\sqrt{n}b_n\right)^{dJ_n}\exp\left(-C\sqrt{n}b_n^{-1/2}\right) \\
= & C\exp\left(C(p+1)\log(n)\right)\exp\left(CdJ_n\log(n)\right)\exp\left(-C\sqrt{n}b_n^{-1/2}\right) \\
\leq & \exp\left(C\left(b_n\log n-\sqrt{n}b_n^{-1/2}\right)\right) \\
\leq & \exp\left(Cb_n\left(\log n-\sqrt{n}b_n^{1/2}\right)\right)\to 0.
\end{aligned}$$

□

Lemma 8

If the conditions of Theorem 4.4 hold, then

$$\hat{\theta}_1-\tilde{\theta}_1 = o_p(1). \qquad \square$$

Proof: Proof will be complete if for positive constants M, L and C

$$P\left(\inf_{||\theta_1-\tilde{\theta}_1||\geq M\,||\theta_2||\leq C\sqrt{b_n}}\sum_{i=1}^{n}\frac{R_i}{\pi_i(\hat{\eta})}Q_i^*(\theta_1,\tilde{\theta}_1,\theta_2)>0\right)\to 1. \tag{7.6}$$

By Lemma 7

$$\sup_{\substack{||\theta_1-\tilde{\theta}_1||\leq M \\ ||\theta_2||\leq C\sqrt{b_n}}} \left| \sum_{i=1}^{n} \frac{R_i}{\pi_i(\hat{\eta})} \left[Q_i^*(\theta_1, \tilde{\theta}_1, \theta_2) - E_s \left[Q_i^*(\theta_1, \tilde{\theta}_1, \theta_2) \right] \right. \right.$$
$$\left. \left. + \tilde{x}_i' \left(\theta_1 - \tilde{\theta}_1 \right) \psi_\tau(\epsilon_i) \right] \right| = o_p(1).$$

Then by Lemma 6

$$\sup_{\substack{||\theta_1-\tilde{\theta}_1||\leq M \\ ||\theta_2||\leq C\sqrt{b_n}}} \left| \sum_{i=1}^{n} \frac{R_i}{\pi_i(\hat{\eta})} \left[\left[Q_i^*(\theta_1, \tilde{\theta}_1, \theta_2) + \tilde{x}_i' \left(\theta_1 - \tilde{\theta}_1 \right) \psi_\tau(\epsilon_i) \right] \right. \right.$$
$$\left. \left. - \frac{1}{2} \left(\theta_1' \Sigma_1 \theta_1 - \tilde{\theta}_1' \Sigma_1 \tilde{\theta}_1 \right) \right] \right| = o_p(1). \tag{7.7}$$

Notice

$$\begin{aligned} \left(\theta_1 - \tilde{\theta}_1 \right)' \sum_{i=1}^{n} \tilde{x}_i \psi_\tau(\epsilon_i) &= \left(\theta_1 - \tilde{\theta}_1 \right)' n^{-1/2} X^{*\prime} \psi_\tau(\epsilon) \\ &= \left(\theta_1 - \tilde{\theta}_1 \right)' \Sigma_1 \tilde{\theta}_1 + o_p(1). \end{aligned} \tag{7.8}$$

Then combining (7.7) and (7.8)

$$\sup_{\substack{||\theta_1-\tilde{\theta}_1||\leq M \\ ||\theta_2||\leq C\sqrt{b_n}}} \left| \sum_{i=1}^{n} \frac{R_i}{\pi_i(\hat{\eta})} \left[\left[Q_i^*(\theta_1, \tilde{\theta}_1, \theta_2) \right] + \left(\theta_1 - \tilde{\theta}_1 \right)' \Sigma_1 \tilde{\theta}_1 \right. \right.$$
$$\left. \left. - \frac{1}{2} \left(\theta_1' \Sigma_1 \theta_1 - \tilde{\theta}_1' \Sigma_1 \tilde{\theta}_1 \right) \right] \right| = o_p(1),$$
$$\Rightarrow \sup_{\substack{||\theta_1-\tilde{\theta}_1||\leq M \\ ||\theta_2||\leq C\sqrt{b_n}}} \left| \sum_{i=1}^{n} \frac{R_i}{\pi_i(\hat{\eta})} \left[\left[Q_i^*(\theta_1, \tilde{\theta}_1, \theta_2) \right] - \frac{1}{2} \left(\theta_1 - \tilde{\theta}_1 \right)' \Sigma_1 \left(\theta_1 - \tilde{\theta}_1 \right) \right] \right| = o_p(1).$$

By condition 5 for any $M > 0$

$$\frac{1}{2}\left(\theta_1 - \tilde{\theta}_1\right)' \Sigma_1 \left(\theta_1 - \tilde{\theta}_1\right) > 0.$$

Thus

$$\lim_{n\to\infty} \inf_{\substack{||\theta_1 - \tilde{\theta}_1|| = M \\ ||\theta_2|| \leq C\sqrt{b_n}}} \left| \sum_{i=1}^{n} \frac{R_i}{\pi_i(\hat{\eta})} Q_i^*(\theta_1, \tilde{\theta}_1, \theta_2) \right| > 0.$$

Then by convexity of Q_i^* and corollary 25 of Eggleston (1958) as $n \to \infty$

$$P\left(\inf_{||\theta_1|| \geq L \; ||\theta_2|| \geq C\sqrt{b_n}} \left|\left| \sum_{i=1}^{n} \frac{R_i}{\pi_i(\hat{\eta})} Q_i^*(\theta_1, \tilde{\theta}_1, \theta_2) \right|\right| > 0 \right).$$

□

7.2 Lemmas for Chapter 5

7.2.1 Definitions

In Chapter 5 the number of potential linear covariates and active linear covariates increases with sample size. That is $X \in \mathbb{R}^{n \times p_n + 1}$ and $X_A \in \mathbb{R}^{n \times q_n + 1}$. With the distinction between active and inactive variables and the high-dimensional nature of the data some of our notation needs to be redefined with dimensions reviewed.

$$
\begin{aligned}
X^* &= \left[1_n \, (I - P_W(B)) X_{A_{-1}}\right] \\
&= (x_1^*, \ldots, x_n^*)' \in \mathbb{R}^{n \times q_n + 1} \\
\Delta_n^B &= \Delta_n' B \Delta_n \in \mathbb{R}^{(q_n+1)\times(q_n+1)}, \\
\Delta_n &= [1_n \, \Delta_{n1} \ldots, \Delta_{np}] \in \mathbb{R}^{n \times q_n + 1} \\
\theta_1 &= \sqrt{n} \, (\beta - \beta_0) \in \mathbb{R}^{q_n+1}, \\
\tilde{x}_i &= n^{-1/2} x_i^* \in \mathbb{R}^{q_n+1}, \\
\tilde{s}_i &= \left(\tilde{x}_i', \tilde{W}(z_i)\right)' \in \mathbb{R}^{q_n + dJ_n + 1}, \\
d_n &= dJ_n + q_n.
\end{aligned}
$$

We continue to use other definitions given in Section 7.1.1, but it is important to note some of the notation has been changed. For instance X^* is now a modification of only the active set variables, which we allow to increase with n.

7.2.2 Technical lemmas for Theorem 5.1

Lemma 9 (x_star_big_q)

If conditions of Theorem 5.1 are satisfied then

(1) $n^{-1/2} X^* = n^{-1/2} \Delta_n + o_p(1)$,

(2) $n^{-1} X^{*\prime} B_n X^* = T_n + o_p(1)$,

and there exists a positive constant C such that

(3)$\lambda_{\max} \left(n^{-1} X^{*\prime} X^*\right) \leq C$. □

Proof: Following definitions and methods provided in proof of Lemma 3 it is sufficient to show

$$n^{-1} ||H - P_W(B) X_{A_{(-1)}}||^2,$$

where $X_{A_{(-1)}} = (I - P_W(B))X_{(-1)}$. Note following proof of 3 and accounting for rates of J_n and q_n as stated in conditions 4 and 12

$$\begin{aligned} n^{-1}||H - PX_A||^2 &\leq n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{q_n}(h_j^*(z_i) - \hat{h}_j(z_i))^2 \\ &= O_p\left(q_n J_n n^{-1}\right) = o(1). \end{aligned}$$

□

Lemma 10

If conditions of Theorem 5.1 are satisfied then for any positive constant L,

$$d_n^{-1}\sup_{||\theta||\leq L}\left|\sum_{i=1}^{n} D_i(\theta, \sqrt{d_n})\right| = o_p(1). \qquad \square$$

Proof: Following Lemma 5.1 of Shi and Li (1995), it can be shown that with probability one $\max_i||\tilde{W}(z_i)|| \leq C_0\sqrt{\frac{dJ_n}{n}}$, for some positive constant C_0. By the definition of $\tilde{x}_i$ and Condition 11, $\max_i||\tilde{x}_i|| \leq C_1\sqrt{\frac{q_n}{n}}$, for some positive constant C_2. Let F_{n1} denote the event $\tilde{s}_{(n)} \leq C_2\sqrt{d_n/n}$, where $\tilde{s}_{(n)} = \max_i||\tilde{s}_i||$ and $C_2 = \max(C_0, C_1)$. From the above analysis, $P(F_{n1}) = 1$. Let F_{n2} denote the event $\max_i |u_{ni}| \leq C_3 J_n^{-r}$, then $P(F_{n2}) = 1$.

Hence, to prove the lemma, it is sufficient to show that $\forall\ \epsilon > 0$,

$$P\left(d_n^{-1}\sup_{||\theta||\leq 1}, \left|\sum_{i=1}^{n} D_i\left(\theta, L\sqrt{d_n}\right)\right| > \epsilon, F_{n1} \cap F_{n2}\right) \to 0. \tag{7.9}$$

Define $\Theta^* \equiv \left\{\theta^* \mid ||\theta^*|| \leq 1,\ \theta \in \mathbb{R}^{d_n+1}\right\}$. We can partition Θ as a union of disjoint regions $\Theta_1, ..., \Theta_{M_n}$, such that the diameter of each region does not exceed $m_0 = \frac{\epsilon}{8C_2L\sqrt{n}}$, where C is a positive constant. This covering can be constructed such that $M_n \leq C\left(\frac{C\sqrt{n}}{\epsilon}\right)^{d_n+1}$. Let $\theta_1^*, ..., \theta_{M_n}^*$ be arbitrary points in $\Theta_1, ..., \Theta_{M_n}$, respectively,

and write $\theta_k^* = (\theta_{k1}^{*\prime}, \theta_{k2}^{*\prime})'$, $k = 1, \ldots, M_n$. Then

$$
\begin{aligned}
& P\left(\sup_{||\theta|| \leq 1} d_n^{-1} \left| \sum_{i=1}^{n} D_i(\theta, L\sqrt{d_n}) \right| > \epsilon, F_{n1} \cap F_{n2} \right) \\
\leq & \sum_{k=1}^{M_n} P\left(\sup_{\theta \in \Theta_k} d_n^{-1} \left| \sum_{i=1}^{n} D_i(\theta, L\sqrt{d_n}) \right| > \epsilon, F_{n1} \cap F_{n2} \right) \\
\leq & \sum_{k=1}^{M_n} P\Bigg(\left| \sum_{i=1}^{n} D_i(\theta_k^*, L\sqrt{d_n}) \right| + \sup_{\theta \in \Theta_k} \left| \sum_{i=1}^{n} \left(D_i(\theta, L\sqrt{d_n}) - D_i(\theta_k^*, L\sqrt{d_n}) \right) \right| \\
& > d_n \epsilon, F_{n1} \cap F_{n2} \Bigg)
\end{aligned}
$$

Let $I(\cdot)$ denote the indicator function, we next show that

$$
\sup_{\theta \in \Theta_k} \left| d_n^{-1} \sum_{i=1}^{n} \left[D_i(\theta, L\sqrt{d_n}) - D_i(\theta_k^*, L\sqrt{d_n}) \right] \right| I\left(F_{n1} \cap F_{n2}\right) < \epsilon/2.
$$

Using (7.3), the triangle inequality, and the earlier derived bounds for $||\tilde{x}_i|||$ and $||\tilde{W}(z_i)||$, we have

$$
\begin{aligned}
& \sup_{\theta\in\Theta_k}\left|d_n^{-1}\sum_{i=1}^{n}\left(D_i(\theta, L\sqrt{d_n}) - D_i(\theta_k^*, L\sqrt{d_n})\right)\right| I\left(F_{n1}\cap F_{n2}\right) \\
= \; & d_n^{-1}\sup_{\theta\in\Theta_k}\left|\sum_{i=1}^{n}\frac{1}{2}\left[\left|\epsilon_i - \tilde{x}_i'\theta_1 L\sqrt{d_n} - \tilde{W}(z_i)'\theta_2 L\sqrt{d_n} - u_{ni}\right| - |\epsilon_i - u_{ni}|\right]\right. \\
& -\sum_{i=1}^{n}\frac{1}{2}E_s\left[\left|\epsilon_i - \tilde{x}_i'\theta_1 L\sqrt{d_n} - \tilde{W}(z_i)'\theta_2 L\sqrt{d_n} - u_{ni}\right| - |\epsilon_i - u_{ni}|\right] \\
& +\sum_{i=1}^{n} L\sqrt{d_n}\left(\tilde{x}_i'\theta_1 + \tilde{W}(z_i)'\theta_2\right)\psi_\tau(\epsilon_i) \\
& -\sum_{i=1}^{n}\frac{1}{2}\left[\left|\epsilon_i - \tilde{x}_i'\theta_{k1}^* L\sqrt{d_n} - \tilde{W}(z_i)'\theta_{k2}^* L\sqrt{d_n} - u_{ni}\right| - |\epsilon_i - u_{ni}|\right] \\
& +\sum_{i=1}^{n}\frac{1}{2}E_s\left[\left|\epsilon_i - \tilde{x}_i'\theta_{k1}^* L\sqrt{d_n} - \tilde{W}(z_i)'\theta_{k2}^* L\sqrt{d_n} - u_{ni}\right| - |\epsilon_i - u_{ni}|\right] \\
& \left.-\sum_{i=1}^{n} L\sqrt{d_n}\left(\tilde{x}_i'\theta_{k1}^* + \tilde{W}(z_i)'\theta_{k2}^*\right)\psi_\tau(\epsilon_i)\right| I\left(F_{n1}\cap F_{n2}\right) \\
\leq \; & 2nLm_0 d_n^{-1/2}\max_i\left[||\tilde{x}_i|| + ||\tilde{W}(z_i)||\right] I\left(F_{n1}\cap F_{n2}\right) \\
\leq \; & 2\sqrt{2}C_2 nLm_0 d_n^{-1/2}\sqrt{d_n/n} = 2\sqrt{2}C_2 L\sqrt{n}m_0 < \epsilon/2,
\end{aligned}
$$

by the definition of m_0.

Therefore, to prove (7.9), we only need to verify

$$
\sum_{k=1}^{M_n} P\left(\left|\sum_{i=1}^{n} D_i(\theta_k^*, L\sqrt{d_n})\right| > d_n\epsilon/2, F_{n1}\cap F_{n2}\right) \to 0. \tag{7.10}
$$

Using methods similar to those in the proof of 4

$$
\max_i\left|D_i(\theta_k^*, L\sqrt{d_n})\right| I\left(F_{n1}\cap F_{n2}\right) \leq C d_n n^{-1/2},
$$

for some positive constant C and

$$\sum_{i=1}^{n} \text{Var}\Big(D_i(\theta) I\left(F_{n1} \cap F_{n2}\right) \big| x_i, z_i \Big) \leq C d_n^2 n^{-1/2}.$$

By Bernstein's inequality, for all n sufficiently large,

$$\begin{aligned} & \sum_{k=1}^{M_n} P\left(\left| \sum_{i=1}^{n} D_i(\theta_k^*, L\sqrt{d_n/n}) \right| > d_n \epsilon/2, F_{n1} \cap F_{n2} \Big| x_i, z_i \right) \\ \leq \; & 2 \sum_{k=1}^{M_n} \exp\left(-C\sqrt{n}\right) \\ \leq \; & C \exp\left(C(d_n+1)\log(n) - C\sqrt{n} \right), \end{aligned}$$

which converges to zero as $n \to \infty$. Note that the upper bound does not depend on $\{x_i, z_i\}$. This implies (7.10). Hence, the proof is complete. □

Lemma 11

If conditions of Theorem 5.1 are satisfied then

$$n^{-1} \sum_{i=1}^{n} (\hat{g}(z_i) - g_0(z_i))^2 = O_p\left(d_n/n\right). \qquad \square$$

Proof: We will first prove that $\forall\ \eta > 0$, there exists an $L > 0$ such that

$$P\left(\inf_{||\theta||=L} d_n^{-1} \sum_{i=1}^{n} (Q_i(\sqrt{d_n}) - Q_i(0)) > 0 \right) \geq 1 - \eta. \tag{7.11}$$

Note that

$$\begin{aligned} d_n^{-1}\sum_{i=1}^{n}(Q_i(\sqrt{d_n})-Q_i(0)) &= d_n^{-1}\sum_{i=1}^{n}D_i(\theta,\sqrt{d_n})+d_n^{-1}\sum_{i=1}^{n}E_s[Q_i(\sqrt{d_n})-Q_i(0)] \\ &\quad -d_n^{-1/2}\sum_{i=1}^{n}(\tilde{x}_i'\theta_1+\tilde{W}(z_i)'\theta_2)\psi_\tau(\epsilon_i) \\ &= G_{n1}+G_{n2}+G_{n3}, \end{aligned}$$

where the definition of G_{ni}, $i = 1, 2, 3$, is clear from the context. By Lemma 10 $G_{n1} = o_p(1)$. Note that $E(G_{n3}) = 0$ and by condition 12

$$\begin{aligned} E(G_{n3}^2) &\leq Cd_n^{-1}\mathrm{E}\left[n^{-1}\theta_1'X_A'(I_n-P)'(I_n-P)X_A\theta_1+||W_B^{-1}||^2\theta_2W'W\theta_2\right] \\ &= O\left(d_n^{-1}||\theta||^2\right). \end{aligned}$$

Therefore $G_{n3} = O_p\left(d_n^{-1/2}||\theta||\right)$. Next, we analyze G_{n2}. Using condition 1 and methods used to analyze D_{n2} from Lemma 5, we have

$$\begin{aligned} G_{n2} &= Cn^{-1}\theta_1'X^{*'}B_nX^*\theta_1(1+o(1))+C\theta_2'W_B^{-1}W_B^2W_B^{-1}C\theta_2(1+o(1)) \\ &\quad +d_n^{-1/2}\sum_{i=1}^{n}f_i(0)u_{ni}\left(\tilde{x}_i'\theta_1+\tilde{W}(z_i)'\theta_2\right), \end{aligned}$$

where the second last inequality follows because $\sum_{i=1} f_i(0)\tilde{x}_i\tilde{W}(z_i) = 0$. By Lemma 9, $n^{-1}\theta_1'X^{*'}B_nX^*\theta_1 = \theta_1'T_n\theta_1 + o_p(1)$. Hence, by condition 11, there exists a finite positive constant c, such that with probability approaching one $n^{-1}\theta_1'X^{*'}B_nX^*\theta_1 + \theta_2W_B^{-1}W_B^2W_B^{-1}\theta_2 \geq c||\theta||^2$. Define $U_n = (u_{n1}, ..., u_{nn})'$. Then, by Schumaker (1981),

$||U_n|| = O_p\left(\sqrt{n}J_n^{-r}\right) = o_p(1)$. Thus, for the linear terms,

$$\begin{aligned} d_n^{-1/2}\sum_{i=1}^{n} f_i(0)u_{ni}\tilde{x}_i'\theta_1 &= d_n^{-1/2}n^{-1/2}\theta_1'X^{*\prime}B_nU_n = o_p(||\theta||), \\ d_n^{-1/2}\sum_{i=1}^{n} f_i(0)u_{ni}\tilde{W}(z_i)'\theta_2 &= d_n^{-1/2}\theta_2'W_B^{-1}W'B_nU_n = o_p(||\theta||). \end{aligned}$$

The above terms are $O_p(||\theta||)$ for the optimal rate of convergence. However, proof still holds since the quadratic term dominates. For L sufficiently large, the always positive quadratic term asymptotically dominates. This proves (7.11).

By convexity, (7.11) implies $||\hat{\theta}|| = O_p(\sqrt{d_n})$, where $\hat{\theta} = (\hat{\theta}_1^T, \hat{\theta}_2^T)^T$. From the definition of $\hat{\theta}$, it follows that $||W_B(\hat{\gamma} - \gamma_0)|| = O_p\left(\sqrt{d_n}\right)$. Using these facts and condition 4,

$$\begin{aligned} n^{-1}\sum_{i=1}^{n} f_i(0)\left(\hat{g}(z_i) - g_0(z_i)\right)^2 &= n^{-1}\sum_{i=1}^{n} f_i(0)\left(W(z_i)'(\hat{\gamma} - \gamma_0) - u_{ni}\right)^2 \\ &\leq n^{-1}\left(\hat{\gamma} - \gamma_0\right)W_B^2\left(\hat{\gamma} - \gamma_0\right) + O_p\left(J_n^{-2r}\right) \\ &= O_p\left(n^{-1}d_n\right). \end{aligned}$$

Then by condition 1, $n^{-1}\sum_{i=1}^{n}\left(\hat{g}(z_i) - g_0(z_i)\right)^2 = O_p\left(n^{-1}d_n\right)$. □

Lemma 12

Let $\tilde{\theta}_1 = \sqrt{n}\left(X^{*\prime}B_nX^*\right)^{-1}X^{*\prime}\psi_\tau(\epsilon)$, where $\psi_\tau(\epsilon) = (\psi_\tau(\epsilon_1), \ldots, \psi_\tau(\epsilon_n))'$. If the conditions of Theorem 5.1 hold then

(1) $||\tilde{\theta}_1|| = O_p\left(\sqrt{q_n}\right)$.

(2) $A_n\Sigma_n^{-1/2}\tilde{\theta}_1 \xrightarrow{d} N(0, G)$, where A_n and Σ_n are defined in Theorem 5.2. □

Proof: (1) The result follows from the observation that, by Lemma 9,

$$\tilde{\theta}_1 = (T_n + o_p(1))^{-1}\Big[n^{-1/2}\Delta_n'\psi_\tau(\epsilon) + n^{-1/2}(H - PX_A)\psi_\tau(\epsilon)\Big],$$

and $n^{-1/2}||H - PX_A|| = o(1)$.

(2)

$$\begin{aligned} A_n\Sigma_n^{-1/2}\tilde{\theta}_1 &= A_n\Sigma_n^{-1/2}T_n^{-1}\left[n^{-1/2}\Delta_n'\psi_\tau(\epsilon)\right](1+o_p(1)) \\ &\quad + A_n\Sigma_n^{-1/2}T_n^{-1}\left[n^{-1/2}(H-PX_A)\right]\psi_\tau(\epsilon)(1+o_p(1)), \end{aligned}$$

where the second term is $o_p(1)$ because $n^{-1/2}||H - PX_A|| = o(1)$. We write

$$A_n\Sigma_n^{-1/2}T_n^{-1}\left[n^{-1/2}\Delta_n'\psi_\tau(\epsilon)\right] = \sum_{i=1}^n D_{ni},$$

where $D_{ni} = n^{-1/2}A_n\Sigma_n^{-1/2}T_n^{-1}\delta_i\psi_\tau(\epsilon_i)$. To verify asymptotic normality, we check the Lindeberg-Feller condition. For any $\epsilon > 0$ and using conditions 1, 11 and 12

$$\begin{aligned} &\sum_{i=1}^n E\left[||D_{ni}||^2 I(||D_{ni}|| > \epsilon)\right] \\ \leq\ & \epsilon^{-2}\sum_{i=1}^n E||D_{ni}||^4 \\ \leq\ & (n\epsilon)^{-2}\sum_{i=1}^n E\left(\psi_\tau^4(\epsilon_i)\left(\delta_i'T_n^{-1}\Sigma_n^{-1/2}A_nA_n^T\Sigma_n^{-1/2}T_n^{-1}\delta_i\right)^2\right) \\ \leq\ & Cn^{-2}\epsilon^{-2}\sum_{i=1}^n E(||\delta_i||^4) = O_p(q_n^2/n) = o_p(1). \end{aligned}$$

The proof is complete by observing that

$$\sum_{i=1}^n E(D_{ni}D_{ni}') = A_n\Sigma_n^{-1/2}T_n^{-1}S_nT_n^{-1}\Sigma_n^{-1/2}A_n \to G.$$

□

The following lemmas will be used to show that $\hat{\theta}_1 - \tilde{\theta}_1 = o_p(1)$. Recall

$$Q_i^*(\theta_1, \tilde{\theta}_1, \theta_2) = \rho_\tau(\epsilon_i - \tilde{x}_i'\theta_1 - \tilde{W}(z_i)'\theta_2 - u_{ni}) - \rho_\tau(\epsilon_i - \tilde{x}_i'\tilde{\theta}_1 - \tilde{W}(z_i)'\theta_2 - u_{ni}).$$

Lemma 13

If the conditions of Theorem 5.1 hold, then

$$\sup_{\substack{||\theta_1 - \tilde{\theta}_1|| \leq M \\ ||\theta_2|| \leq C\sqrt{d_n}}} \left| \sum_{i=1}^n E_s \left[Q_i^*(\theta_1, \tilde{\theta}_1, \theta_2) \right] - \frac{1}{2} \left[\theta_1' T_n \theta_1 - \tilde{\theta}_1' T_n \tilde{\theta}_1 \right] (1 + o(1)) \right| = o_p(1). \quad \square$$

Proof: Using methods from Lemma 6 and results from Lemma 9 we have

$$\begin{aligned} & \sum_{i=1}^n E_s \left[Q_i^*(\theta_1, \tilde{\theta}_1, \theta_2) \right] \\ = \; & \frac{1}{2} \left[\theta_1' T_n \theta_1 - \tilde{\theta}_1' T_n \tilde{\theta}_1 \right] (1 + o(1)) + \frac{1}{2} n^{-1/2} (\theta_1 - \tilde{\theta}_1)' X^{*'} B_n U_n (1 + o(1)). \end{aligned}$$

The proof is complete by noting that $\sup_{||\theta_1 - \tilde{\theta}_1|| \leq M, \, ||\theta_2|| \leq C\sqrt{d_n}} |n^{-1/2} (\theta_1 - \tilde{\theta}_1)' X^{*'} B_n U_n (1 + o(1))| = o_p(1)$. $\square$

Lemma 14

If the conditions of Theorem 5.1 hold, then for any given positive constants M and C,

$$\sup_{\substack{||\theta_1 - \tilde{\theta}_1|| \leq M \\ ||\theta_2|| \leq C\sqrt{d_n}}} \left| \sum_{i=1}^n \left[Q_i^*(\theta_1, \tilde{\theta}_1, \theta_2) - E_s \left[Q_i^*(\theta_1, \tilde{\theta}_1, \theta_2) \right] + \tilde{x}_i' \left(\theta_1 - \tilde{\theta}_1 \right) \psi_\tau(\epsilon_i) \right] \right| = o_p(1). \quad \square$$

Proof: Define $A_i(\theta_1, \tilde{\theta}_1, \theta_2) = Q_i^*(\theta_1, \tilde{\theta}_1, \theta_2) - E_s\left[Q_i^*(\theta_1, \tilde{\theta}_1, \theta_2)\right] + \tilde{x}_i'\left(\theta_1 - \tilde{\theta}_1\right)\psi_\tau(\epsilon_i)$. We note that

$$\begin{aligned} A_i(\theta_1, \tilde{\theta}_1, \theta_2) &= \frac{1}{2}\left[|\epsilon_i - \tilde{x}_i'\theta_1 - \tilde{W}(z_i)'\theta_2 - u_{ni}| - |\epsilon_i - \tilde{x}_i'\tilde{\theta}_1 - \tilde{W}(z_i)'\theta_2 - u_{ni}|\right] \\ &+(\tau - 1/2)(\tilde{x}_i'(\tilde{\theta}_1 - \theta_1)) \\ &-\frac{1}{2}E_s\left[|\epsilon_i - \tilde{x}_i'\theta_1 - \tilde{W}(z_i)'\theta_2 - u_{ni}| - |\epsilon_i - \tilde{x}_i'\tilde{\theta}_1 - \tilde{W}(z_i)'\theta_2 - u_{ni}|\right] \\ &-E_s\left[(\tau - 1/2)(\tilde{x}_i'(\tilde{\theta}_1 - \theta_1))\right] \\ &+\tilde{x}_i'(\theta_1 - \tilde{\theta}_1)\psi_\tau(\epsilon_i). \end{aligned}$$

Let F_{n1} and F_{n2} be defined as in Lemma 10. Then proof will be complete if we verify

$$P\left(\sup_{||\theta_1 - \tilde{\theta}_1|| \le M,\ ||\theta_2|| \le C\sqrt{d_n}} \left|\sum_{i=1}^n A_i(\theta_1, \tilde{\theta}_1, \theta_2)\right|,\ F_{n1} \cap F_{n2}\right) \to 0.$$

Similarly as in the proof of Lemma 10, let $\Theta_1 = \left\{\theta_1 :\ ||\theta - \tilde{\theta}_1|| \le M,\ \theta_1 \in \mathbb{R}^{q_n+1}\right\}$ and $\Theta_2 = \left\{\theta_2 :\ ||\theta_2|| \le C\sqrt{d_n},\ \theta_2 \in \mathbb{R}^{dJ_n}\right\}$. We can partition Θ_1 (similarly Θ_2), into disjoint regions $\Theta_{11}, ..., \Theta_{1K_n}$ $(\Theta_{21}, ..., \Theta_{2L_n})$ such that the diameter of each region does not exceed $m_0^* = \frac{C\epsilon}{2\sqrt{nd_n}}$. These partitions can be constructed such that $K_n \le C\left(\frac{C\sqrt{nd_n}}{2\epsilon}\right)^{q_n+1}$ and $L_n \le C\left(\frac{C\sqrt{n}d_n}{2\epsilon}\right)^{dJ_n}$. Let $\theta_{11}^*, ..., \theta_{1K_n}^*$ be arbitrary points in $\Theta_{11}, ..., \Theta_{1K_n}$, respectively; similarly, let $\theta_{21}^*, ..., \theta_{2L_n}^*$ be arbitrary points in $\Theta_{21}, ..., \Theta_{2L_n}$, respectively.

Then the left side of (7.12) is bounded by

$$\sum_{l=1}^{L_n}\sum_{k=1}^{K_n} P\left(\sup_{\theta_1 \in \Theta_{1k}\ \theta_2 \in \Theta_{2l}} \left|\sum_{i=1}^n A_i(\theta_{1k}, \tilde{\theta}_1, \theta_{2l}) + A_i(\theta_1, \tilde{\theta}_1, \theta_2) - A_i(\theta_{1k}, \tilde{\theta}_1, \theta_{2l})\right| > \epsilon \right| \tilde{s}_{(n)} < C\sqrt{d_n/n}\right).$$

Following steps shown in Lemma 7

$$\begin{aligned} & Q_i^*(\theta_1, \tilde{\theta}_1, \theta_2) - Q_i^*(\theta_{1k}, \tilde{\theta}_1, \theta_{2l}) \\ \leq \; & 2\max_i ||\tilde{s}_i|| \sup_{\theta_1 \in \Theta_{1k}\, \theta_2 \in \Theta_{2l}} \left[||\theta_1 - \theta_{1k}|| + ||\theta_2 - \theta_{2l}|| \right]. \end{aligned}$$

Using the above inequality, the definition of $A_i(\theta_1, \tilde{\theta}_1, \theta_2)$, m_0^* and condition 11

$$\begin{aligned} & \sup_{\theta_1 \in \Theta_{1k}\, \theta_2 \in \Theta_{2l}} \sum_{i=1}^{n} \left| A_i(\theta_1, \tilde{\theta}_1, \theta_2) - A_i(\bar{\theta}_{l_1}, \tilde{\theta}_1, \bar{\theta}_{l_2}) \right| I(\tilde{s}_{(n)} \leq C\sqrt{d_n/n}) \\ \leq \; & 5n \max_i ||\tilde{s}_i|| \sup_{\theta_1 \in \Theta_{1k}\, \theta_2 \in \Theta_{2l}} \left[||\theta_1 - \theta_{1k}|| + ||\theta_2 - \theta_{2l}|| \right] I(\tilde{s}_{(n)} \leq C\sqrt{d_n/n}) \\ \leq \; & C m_o^* \sqrt{n d_n} \end{aligned}$$

By definition of m_0^* and condition 11

$$\begin{aligned} \sum_{i=1}^{n} \left| A_i(\theta_1, \tilde{\theta}_1, \theta_2) - A_i(\theta_{1k}, \tilde{\theta}_1, \theta_{2l}) \right| \; & \leq \; 2 \sum_{i=1}^{n} \max_i |\tilde{x}_i'(\theta_1 - \theta_{1k}) + \tilde{W}(z_i)'(\theta_2 - \theta_{2l})| \\ & \leq \; 2C\sqrt{n(J_n + q_n)} m_0^* < \epsilon/2. \end{aligned}$$

Proof will be complete if

$$\sum_{l=1}^{L_n} \sum_{k=1}^{K_n} P\left(\left| \sum_{i=1}^{n} A_i(\theta_{1k}, \tilde{\theta}_1, \theta_{2l}) \right| > \epsilon/2 \right).$$

Bernstein's inequality will be used and the variance and maximum of $A_i(\theta_{1k}, \tilde{\theta}_1, \theta_{2l})$ is needed. Assuming $\tilde{s}_{(n)} < C\sqrt{d_n/n}$ and noting this depends on $\max_i ||\tilde{x}_i|| < \sqrt{\frac{q_n}{n}}$. Also noting that this lemma requires that there exists a positive constant C such that $||\theta_1 - \tilde{\theta}_1|| < C$ then, following steps used in Lemma 7, the maximum has the following

upper bound

$$\max_i \left| A_i(\theta_{1k}, \tilde{\theta}_1, \theta_{2l}) \right| \leq 3 \max_i ||\tilde{x}_i|| ||\theta_1 - \tilde{\theta}_1|| \leq C\sqrt{q_n/n}.$$

Using Knight's identity (Knight 1998)

$$\begin{aligned}
& \rho_\tau(\epsilon_i - \tilde{x}_i'\theta_{1k} - \tilde{W}(z_i)'\theta_{2l} - u_{ni}) - \rho_\tau(\epsilon_i - \tilde{x}_i'\tilde{\theta}_1 - \tilde{W}(z_i)'\theta_{2l} - u_{ni}) \\
& +(\theta_{1k} - \tilde{\theta}_1)'\tilde{x}_i\psi_\tau(\epsilon_i) \\
= \; & (\theta_{1k} - \tilde{\theta}_1)'\tilde{x}_i \left(I(\epsilon_i < \tilde{x}_i'\tilde{\theta}_1 + \tilde{W}(z_i)'\theta_{2l} + u_{ni}) - I(\epsilon_i < 0) \right) \\
+ \; & \int_0^{\tilde{x}_i'(\theta_{1k} - \tilde{\theta}_1)} I(\epsilon_i \leq s + \tilde{x}_i'\tilde{\theta}_1 + \tilde{W}(z_i)'\theta_{2l} + u_{ni}) - I(\epsilon_i \leq \tilde{x}_i'\tilde{\theta}_1 + \tilde{W}(z_i)'\theta_{2l} + u_{ni}) ds \\
= \; & A_{i1} + A_{i2}
\end{aligned}$$

To get an upper bound for $\sum_{i=1}^n \text{Var}(A_i(\theta_{1k}, \tilde{\theta}_1, \theta_{2l}))$ we analyze $\sum_{i=1}^n E\left[A_{i1}^2\right]$ and $\sum_{i=1}^n E\left[A_{i2}^2\right]$. Using Lemma 12, conditions 1 and 12, the definitions of θ_{1k} and θ_{2l}, the rate of $\max_i |u_{ni}|$ and $\max_i ||\tilde{s}_i|| < \sqrt{d_n/n}$. Then using methods from Lemma 7 to evaluate D_{i1}^2

$$\begin{aligned}
\sum_{i=1}^n E\left[A_{i1}^2\right] &\leq C \max_i \left| \tilde{x}_i'\tilde{\theta}_1 + \tilde{W}(z_i)'\theta_{2l} + u_{ni} \right| \sum_{i=1}^n E\left[\left(\tilde{x}_i' \left(\theta_{1k} - \tilde{\theta}_1 \right) \right)^2 \right] \\
&\leq C \left(\sqrt{d_n} n^{-1/2} + J_n^{-r} \right) E\left[(\theta_{1k} - \tilde{\theta}_1)' \frac{1}{n} \sum_{i=1}^n x_i^* x_i^{*\prime} (\theta_{1k} - \tilde{\theta}_1) \right] \\
&\leq C\sqrt{d_n} n^{-1/2}.
\end{aligned}$$

Using similar techniques for D_{i2} from the proof of Lemma 7

$$\begin{aligned}
& \sum_{i=1}^{n} E\left[A_{i2}^2\right] \\
\leq\ & \max_i \left|\tilde{x}_i'(\theta_{1k} - \tilde{\theta}_1)\right| \\
\times\ & \sum_{i=1}^{n} E\left[\int_0^{\tilde{x}_i'(\theta_{1k}-\tilde{\theta}_1)} \left[F_i(s + \tilde{x}_i'\tilde{\theta}_1 + \tilde{W}(z_i)'\theta_{2l} + u_{ni})\right.\right. \\
& \left.\left. - F_i(\tilde{x}_i'\tilde{\theta}_1 + \tilde{W}(z_i)'\theta_{2l} + u_{ni})ds\right]\right] \\
\leq\ & \sqrt{q_n} C n^{-1/2} \left[\theta_1' \frac{1}{n} \sum_{i=1}^{n} E\left[x_i^* x_i^{*\prime}\right] \theta_1\right] \\
\leq\ & \sqrt{q_n} C n^{-1/2}.
\end{aligned}$$

Therefore

$$\sum_{i=1}^{n} \mathrm{Var}(A_i(\theta_{1k}, \tilde{\theta}_1, \theta_{2l})) I\left(\tilde{s}_{(n)} < C\sqrt{d_n/n}\right) \leq C\sqrt{\frac{d_n}{n}}.$$

Using Bernstein's inequality and conditions 4 and 12

$$\begin{aligned}
& \sum_{l=1}^{L_n} \sum_{k=1}^{K_n} P\left(\left|\sum_{i=1}^{n} A_i(\theta_{1k}, \tilde{\theta}_1, \theta_{2l})\right| > \epsilon/2 \middle| \tilde{s}_{(n)} < C\sqrt{d_n/n}\right) \\
\leq\ & \sum_{l=1}^{L_n} \sum_{k=1}^{K_n} \exp\left(\frac{-\epsilon^2/4}{C\sqrt{d_n} n^{-1/2} + \epsilon C\sqrt{q_n} n^{-1/2}}\right) \\
\leq\ & \sum_{l=1}^{L_n} \sum_{k=1}^{K_n} \exp\left(-C\sqrt{n} d_n^{-1/2}\right) \\
\leq\ & C\left(C\sqrt{n d_n}\right)^{q_n+1} \left(C\sqrt{n} d_n\right)^{dJ_n} \exp\left(-C\sqrt{n} d_n^{-1/2}\right) \\
\leq\ & \exp\left(C d_n \left(\log n - \sqrt{n} d_n^{1/2}\right)\right) \to 0.
\end{aligned}$$

□

Lemma 15

If conditions 1-12 hold and $q_n = O(n^{c_1})$ with $c_1 < 1/2$ then

$$\hat{\theta}_1 - \tilde{\theta}_1 = o_p(1). \qquad \square$$

Proof: Proof will be complete if for positive constants M, L and C

$$P\left(\inf_{||\theta_1 - \tilde{\theta}_1|| \geq M \, ||\theta_2|| \leq C\sqrt{d_n}} \sum_{i=1}^{n} Q_i^*(\theta_1, \tilde{\theta}_1, \theta_2) > 0\right) \to 1. \tag{7.12}$$

By Lemma 14

$$\sup_{\substack{||\theta_1 - \tilde{\theta}_1|| \leq M \\ ||\theta_2|| \leq C\sqrt{d_n}}} \left| \sum_{i=1}^{n} Q_i^*(\theta_1, \tilde{\theta}_1, \theta_2) - E_s\left[Q_i^*(\theta_1, \tilde{\theta}_1, \theta_2)\right] + \tilde{x}_i'\left(\theta_1 - \tilde{\theta}_1\right)\psi_\tau(\epsilon_i) \right| = o_p(1).$$

Then by Lemma 13

$$\sup_{\substack{||\theta_1 - \tilde{\theta}_1|| \leq M \\ ||\theta_2|| \leq C\sqrt{d_n}}} \left| \sum_{i=1}^{n} \left[Q_i^*(\theta_1, \tilde{\theta}_1, \theta_2) + \tilde{x}_i'\left(\theta_1 - \tilde{\theta}_1\right)\psi_\tau(\epsilon_i)\right] - \frac{1}{2}\left(\theta_1' T_n \theta_1 - \tilde{\theta}_1' T_n \tilde{\theta}_1\right) \right| = o_p(1). \tag{7.13}$$

Notice

$$\begin{aligned} \left(\theta_1 - \tilde{\theta}_1\right)' \sum_{i=1}^{n} \tilde{x}_i \psi_\tau(\epsilon_i) &= \left(\theta_1 - \tilde{\theta}_1\right)' n^{-1/2} X^{*\prime} \psi_\tau(\epsilon) \\ &= \left(\theta_1 - \tilde{\theta}_1\right)' T_n \tilde{\theta}_1 + o_p(1). \end{aligned} \tag{7.14}$$

Then combining (7.13) and (7.14)

$$\sup_{\substack{||\theta_1-\tilde{\theta}_1||\leq M \\ ||\theta_2||\leq C\sqrt{d_n}}} \left| \sum_{i=1}^{n} \left[Q_i^*(\theta_1, \tilde{\theta}_1, \theta_2) \right] + \left(\theta_1 - \tilde{\theta}_1\right)' T_n \tilde{\theta}_1 - \tfrac{1}{2} \left(\theta_1' T_n \theta_1 - \tilde{\theta}_1' T_n \tilde{\theta}_1 \right) \right| = o_p(1),$$

$$\Rightarrow \sup_{\substack{||\theta_1-\tilde{\theta}_1||\leq M \\ ||\theta_2||\leq C\sqrt{d_n}}} \left| \sum_{i=1}^{n} \left[Q_i^*(\theta_1, \tilde{\theta}_1, \theta_2) \right] - \tfrac{1}{2} \left(\theta_1 - \tilde{\theta}_1\right)' T_n \left(\theta_1 - \tilde{\theta}_1\right) \right| = o_p(1).$$

By conditions 1 and 11 for any $M > 0$

$$\frac{1}{2} \left(\theta_1 - \tilde{\theta}_1\right)' T_n \left(\theta_1 - \tilde{\theta}_1\right) > 0.$$

Thus

$$\lim_{n\to\infty} \inf_{\substack{||\theta_1-\tilde{\theta}_1||=M \\ ||\theta_2||\leq C\sqrt{d_n}}} \left| \sum_{i=1}^{n} Q_i^*(\theta_1, \tilde{\theta}_1, \theta_2) \right| > 0.$$

Then by convexity of Q_i^* and corollary 25 of Eggleston (1958) as $n \to \infty$

$$P\left(\inf_{||\theta_1||\geq L \; ||\theta_2||\geq C\sqrt{d_n}} \left|\left| \sum_{i=1}^{n} Q_i^*(\theta_1, \tilde{\theta}_1, \theta_2) \right|\right| > 0 \right).$$

□

7.2.3 Technical lemmas for Theorem 5.3

Lemma 16

Consider the function $k(x) - l(x)$ where both k and l are convex with subdifferential functions $\partial k(x)$ and $\partial l(x)$. Let x^* be a point that has neighborhood U such that $\partial l(x) \cap \partial k(x^*) \neq \emptyset, \forall\, x \in U \cap \text{dom}(k)$. Then x^* is a local minimizer of $k(x) - l(x)$. □

Proof: Proofs available in Tao and An (1997). □

Lemma 17

Assume the conditions of Theorem 5.3 hold and $\log(p_n) = o(n\lambda^2)$ and $n\lambda^2 \to \infty$ then

$$P\left(\max_{q_n+1\leq j\leq p_n} \frac{1}{n}\left|\sum_{i=1}^{n} x_{ij}\left[I(Y_i - x'_{Ai}\beta_{\mathbf{01}} - g_0(z_i) \leq 0) - \tau\right]\right| > \lambda/2\right) \to 0.$$

Proof follows directly from Lemma 4.2 from Wang et al. (2012). □

Lemma 18

Assume the conditions of Theorem 5.3 hold, $n\lambda^2 \to \infty$, $b_n \log(n) = o(n\lambda)$ and $\log p_n = o(n\lambda^2)$. Then for some positive constant C,

$$P\Bigg(\max_{q_n+1\leq j\leq p_n} \sup_{\substack{||\beta-\beta_0||\leq C\sqrt{\frac{q_n}{n}} \\ ||\gamma-\gamma_0||\leq C\sqrt{\frac{d_n}{n}}}} \Bigg|\sum_{i=1}^{n} x_{ij}\Big[I(Y_i - x'_{Ai}\beta - W(z_i)'\gamma \leq 0) - I(Y_i - x'_{Ai}\beta_{\mathbf{01}} - g_0(z_i) \leq 0) - P(Y_i - x'_{Ai}\beta - W(z_i)'\gamma \leq 0) + P(Y_i - x'_{Ai}\beta_{\mathbf{01}} - g_0(z_i) \leq 0)\Big]\Bigg| > n\lambda\Bigg) \to 0 \ \ \forall\ j.$$

Proof: Using the approach of Welsh (1989) we consider the sets $\mathcal{B} = \left\{\beta : ||\beta - \beta_0|| \leq C\sqrt{\frac{q_n}{n}}\right\}$ and $\mathcal{G} = \left\{\gamma : ||\gamma - \gamma_0|| \leq C\sqrt{\frac{d_n}{n}}\right\}$. The set of $\mathcal{B}$ and $\mathcal{G}$ can be covered with with a net of balls with radius $C\sqrt{q_n/n^5}$ and $C\sqrt{d_n/n^5}$ respectively and for some constant $C > 0$ with cardinality $N_1 \equiv |\mathcal{B}| \leq Cn^{4q_n}$ and $N_2 \equiv |\mathcal{G}| \leq Cn^{4d_n}$. Denote the N_1 balls by $\beta(t_1), ..., \beta(t_{N_1})$, where the ball $\beta(t_k)$ is centered at t_k, $k = 1, ..., N_1$ and use similar notation for the balls $\gamma(u_1), ..., \gamma(u_{N_2})$. For ease of notation define $\epsilon_i(\beta, \gamma) = Y_i - x_{Ai}'\beta - W(z_i)'\gamma$ and $\epsilon_i = Y_i - x_{Ai}'\beta_{\mathbf{01}} - g_0(z_i)$.

$$
\begin{aligned}
& P\Bigg(\sup_{\substack{||\beta-\beta_{\mathbf{01}}||\le C\sqrt{\frac{q_n}{n}} \\ ||\gamma-\gamma_0||\le C\sqrt{\frac{d_n}{n}}}} \Bigg| \sum_{i=1}^{n} x_{ij} \Big[I(\epsilon_i(\beta,\gamma)\le 0) - I(\epsilon_i \le 0) - P(\epsilon_i(\beta,\gamma)\le 0) \\
& + P(\epsilon_i \le 0) \Big] \Bigg| \Bigg) \\
\le & \sum_{l=1}^{N_2}\sum_{k=1}^{N_1} P\Bigg(\sup_{\substack{||\tilde{\beta}-t_k||\le C\sqrt{q_n/n^5} \\ ||\tilde{\gamma}-u_l||\le C\sqrt{d_n/n^5}}} \Bigg| \sum_{i=1}^{n} x_{ij} \Big[I(\epsilon_i(\tilde{\beta},\tilde{\gamma})\le 0) - I(\epsilon_i \le 0) \\
& - P\left(\epsilon_i(\tilde{\beta},\tilde{\gamma})\le 0\right) + P\left(\epsilon_i \le 0\right) \Big] \Bigg| > n\lambda \Bigg) \\
= & \sum_{l=1}^{N_2}\sum_{k=1}^{N_1} P\Bigg(\sup_{\substack{||\tilde{\beta}-t_k||\le C\sqrt{q_n/n^5} \\ ||\tilde{\gamma}-u_l||\le C\sqrt{d_n/n^5}}} \Bigg| \sum_{i=1}^{n} x_{ij} \Big[I(\epsilon_i(\tilde{\beta},\tilde{\gamma})\le 0) - I(\epsilon_i \le 0) \\
& + I(\epsilon_i(t_k,u_l)\le 0) - I(\epsilon_i(t_k,u_l)\le 0) - P\left(\epsilon_i(\tilde{\beta},\tilde{\gamma})\le 0\right) + P\left(\epsilon_i \le 0\right) \\
& - P\left(\epsilon_i(t_k,u_l)\le 0\right) + P\left(\epsilon_i(t_k,u_l)\le 0\right) \le 0 \Big] \Bigg| > n\lambda \Bigg) \\
\le & \sum_{l=1}^{N_2}\sum_{k=1}^{N_1} P\Bigg(\Bigg| \sum_{i=1}^{n} x_{ij} \Big[I(\epsilon_i(t_k,u_l)\le 0) - I(\epsilon_i \le 0) - P\left(\epsilon_i(t_k,u_l)\le 0\right) \\
& + P\left(\epsilon_i \le 0\right) \Big] \Bigg| > n\lambda/2 \Bigg) \\
& + \sum_{l=1}^{N_2}\sum_{k=1}^{N_1} P\Bigg(\sup_{\substack{||\tilde{\beta}-t_k||\le C\sqrt{q_n/n^5} \\ ||\tilde{\gamma}-u_l||\le C\sqrt{d_n/n^5}}} \Bigg| \sum_{i=1}^{n} x_{ij} \Big[I\left(\epsilon_i(\tilde{\beta},\tilde{\gamma})\le 0\right) - I\left(\epsilon_i(t_k,u_l)\le 0\right) \\
& - P\left(\epsilon_i(\tilde{\beta},\tilde{\gamma})\le 0\right) + P\left(\epsilon_i(t_k,u_l)\le 0\right) \Big] \Bigg| > n\lambda/2 \Bigg) \\
\equiv & \; I_{nj1} + I_{nj2}.
\end{aligned}
$$

First we will evaluate I_{nj1} using Bernstein's inequality. Define

$$\xi_{ij} = x_{ij}\left[I(\epsilon_i(t_k,u_l)\leq 0) - I(\epsilon_i\leq 0) - P\left(\epsilon_i(t_k,u_l)\leq 0\right) + P\left(\epsilon_i\leq 0\right)\right],$$

which are bounded, independent mean-zero random variables. For the variance

$$\begin{aligned}
Var\left(\xi_{ij}\right) &= E\left[x_{ij}^2\left(I(\epsilon_i(t_k,u_l)\leq 0) - I(\epsilon_i\leq 0) - P\left(\epsilon_i(t_k,u_l)\leq 0\right) + P\left(\epsilon_i\leq 0\right)\right)^2\right] \\
&= E\Bigg[x_{ij}^2\Big(\left(I(\epsilon_i(t_k,u_l)\leq 0) - P\left(\epsilon_i(t_k,u_l)\leq 0\right)\right)^2 + \left(I(\epsilon_i\leq 0) - P\left(\epsilon_i\leq 0\right)\right)^2 \\
&\quad -2\left(I(\epsilon_i(t_k,u_l)\leq 0) - P\left(\epsilon_i(t_k,u_l)\leq 0\right)\right)\left(I(\epsilon_i\leq 0) - P\left(\epsilon_i\leq 0\right)\right)\Big)\Bigg] \\
&= E\Big[x_{ij}^2\Big(F_i({x_{Ai}}'(t_k-\beta_{\mathbf{01}}) + W(z_i)'(u_l-\gamma_0) - u_{ni}) \\
&\quad \times(1 - F_i({x_{Ai}}'(t_k-\beta_{\mathbf{01}}) + W(z_i)'(u_l-\gamma_0) - u_{ni}) + F_i(0)(1-F_i(0)) \\
&\quad +F_i(0)F_i({x_{Ai}}'(t_k-\beta_{\mathbf{01}}) + W(z_i)'(u_l-\gamma_0) - u_{ni}) \\
&\quad \times\Big(2 - F_i(\min({x_{Ai}}'(t_k-\beta_{\mathbf{01}}) + W(z_i)'(u_l-\gamma_0) - u_{ni}, 0))\Big)\Big)\Big] \\
&\leq CE\left[{x_{Ai}}'(t_k-\beta_{\mathbf{01}}) + W(z_i)'(u_l-\gamma_0) - u_{ni}\right] \\
&\leq \sup_i\left(||x_{Ai}||\cdot||t_k-\beta_{\mathbf{01}}|| + ||W(z_i)||\cdot||u_l-\gamma_0|| + ||u_{ni}||\right).
\end{aligned}$$

The min term comes from

$$\begin{aligned}
E\left[I(\epsilon_i(t_k,u_l)\leq 0)I(\epsilon_i\leq 0)\right] &= P\left(\epsilon_i(t_k,u_l)\leq 0, \epsilon_i\leq 0\right) \\
&= F_i(\min({x_{Ai}}'(t_k-\beta_{\mathbf{01}}) + W(z_i)'(u_l-\gamma_0) - u_{ni}, 0)).
\end{aligned}$$

Therefore

$$\begin{aligned}
&\textstyle\sum_{i=1}^n \mathrm{Var}(\xi_{ij}) \leq Cn\sup_i\left(||x_{Ai}||\cdot||t_k-\beta_{\mathbf{01}}|| + ||W(z_i)||\cdot||u_l-\gamma_0|| + ||u_{ni}||\right) \\
\leq\ & Cn\left(\sqrt{q_n}\sqrt{q_n/n} + \sqrt{dJ_n}\sqrt{d_n/n} + (dJ_n)^{-r}\right) \leq Cd_n\sqrt{n}.
\end{aligned}$$

Then using Bernstein's inequality

$$\begin{aligned} I_{nj1} &\leq N_1 N_2 \exp\left(-\frac{n^2\lambda^2/8}{d_n\sqrt{n}+(1/3)2n\lambda}\right) \\ &\leq N_1 N_2 \exp\left(-\frac{n^2\lambda^2}{C\left(d_n\sqrt{n}+n\lambda\right)}\right) \\ &\leq N_1 N_2 \exp\left(-Cn\lambda\right) \\ &\leq Cn^{4q_n} n^{4d_n} \exp\left(-Cn\lambda\right) \\ &= Cn^{8q_n+4dJ_n} \exp\left(-Cn\lambda\right) \\ &= C\exp\left(\left(8q_n+4dJ_n\right)\log(n)-Cn\lambda\right). \end{aligned}$$

For I_{nj2} note that the function $I(x \leq s)$ is an increasing function in s and

$$\begin{aligned} I(\epsilon_i(\tilde{\beta},\tilde{\gamma}) \leq 0) &= I\left(Y_i - x_{Ai}{}'t_k - W(z_i)'u_l - x_{Ai}{}'(\tilde{\beta}-t_k) - W(z_i)'(\tilde{\gamma}-u_l) \leq 0\right) \\ &= I\left(\epsilon_i(t_k,u_l) \leq x_{Ai}{}'(\tilde{\beta}-t_k) + W(z_i)'(\tilde{\gamma}-u_l)\right). \end{aligned}$$

Therefore

$$\begin{aligned}
&\sup_{\substack{||\tilde{\beta}_1-t_k||\leq C\sqrt{q_n/n^5}\\ ||\tilde{\gamma}-u_l||\leq C\sqrt{d_n/n^5}}}\left|\sum_{i=1}^{n} x_{ij}\Big[I\left(\epsilon_i\left(\tilde{\beta}_1,\tilde{\gamma}\right)\leq 0\right)-I\left(\epsilon_i\left(t_k,u_l\right)\leq 0\right)\right.\\
&\left.-P\left(\epsilon_i\left(\tilde{\beta}_1,\tilde{\gamma}\right)\leq 0\right)+P\left(\epsilon_i\left(t_k,u_l\right)\leq 0\right)\Big]\right|\\
\leq\ &\sum_{i=1}^{n}|x_{ij}|\Bigg[I\left(\epsilon_i\left(t_k,u_l\right)\leq C\sqrt{q_n/n^5}||x_{Ai}||+C||W(z_i)||\sqrt{d_n/n^5}\right)\\
&-I\left(\epsilon_i\left(t_k,u_l\right)\leq 0\right)\\
&-P\left(\epsilon_i\left(t_k,u_l\right)\leq -C\sqrt{q_n/n^5}||x_{Ai}||-C\sqrt{d_n/n^5}||W(z_i)||\right)+P\left(\epsilon_i\left(t_k,u_l\right)\leq 0\right)\Bigg]\\
=\ &\sum_{i=1}^{n}|x_{ij}|\Bigg[I\left(\epsilon_i\left(t_k,u_l\right)\leq C\sqrt{q_n/n^5}||x_{Ai}||+C||W(z_i)||\sqrt{d_n/n^5}\right)\\
&-I\left(\epsilon_i\left(t_k,u_l\right)\leq 0\right)\\
&-P\left(\epsilon_i\left(t_k,u_l\right)\leq C\sqrt{q_n/n^5}||x_{Ai}||+C||W(z_i)||\sqrt{d_n/n^5}\right)+P\left(\epsilon_i\left(t_k,u_l\right)\leq 0\right)\Bigg]\\
&+\sum_{i=1}^{n}|x_{ij}|\Bigg[P\left(\epsilon_i\left(t_k,u_l\right)\leq C\sqrt{q_n/n^5}||x_{Ai}||+C||W(z_i)||\sqrt{d_n/n^5}\right)\\
&-P\left(\epsilon_i\left(t_k,u_l\right)\leq -C\sqrt{q_n/n^5}||x_{Ai}||-C\sqrt{d_n/n^5}||W(z_i)||\right)\Bigg].
\end{aligned}$$

For the second sum

$$\begin{aligned}
&\sum_{i=1}^{n}|x_{ij}|\Bigg[P\left(\epsilon_i\left(t_k,u_l\right)\leq C\sqrt{q_n/n^5}||x_{Ai}||+C||W(z_i)||\sqrt{d_n/n^5}\right)\\
&-P\left(\epsilon_i\left(t_k,u_l\right)\leq -C\sqrt{q_n/n^5}||x_{Ai}||-C\sqrt{d_n/n^5}||W(z_i)||\right)\Bigg]\\
\leq\ &C\sum_{i=1}^{n}|x_{ij}|\sqrt{q_n/n^5}||x_{Ai}||+C||W(z_i)||\sqrt{d_n/n^5}\leq CdJ_n\sqrt{d_n}n^{-3/2}=o(1).
\end{aligned}$$

To show $I_{nj2} \to 0$ it will be sufficient to show

$$\sum_{k=1}^{N} P\left(\sum_{i=1}^{n} |x_{ij}| \left[I\left(\epsilon_i\left(t_k, u_l\right) \leq C\sqrt{q_n/n^5}||x_{Ai}|| + C||W(z_i)||\sqrt{d_n/n^5}\right)\right.\right.$$
$$-I\left(\epsilon_i\left(t_k, u_l\right) \leq 0\right) - P\left(\epsilon_i\left(t_k, u_l\right) \leq C\sqrt{q_n/n^5}||x_{Ai}|| + C||W(z_i)||\sqrt{d_n/n^5}\right)$$
$$\left.\left. +P\left(\epsilon_i\left(t_k, u_l\right) \leq 0\right)\right] \geq \frac{n\lambda}{4}\right) \to 0.$$

Define

$$\begin{aligned}\alpha_{ij} &= |x_{ij}|\left[I\left(\epsilon_i\left(t_k, u_l\right) \leq C\sqrt{q_n/n^5}||x_{Ai}|| + C||W(z_i)||\sqrt{d_n/n^5}\right) - I\left(\epsilon_i\left(t_k, u_l\right) \leq 0\right)\right. \\ &- \left. P\left(\epsilon_i\left(t_k, u_l\right) \leq C\sqrt{q_n/n^5}||x_{Ai}|| + C||W(z_i)||\sqrt{d_n/n^5}\right) + P\left(\epsilon_i\left(t_k, u_l\right) \leq 0\right)\right],\end{aligned}$$

then by condition 11 we have a sum of bounded random variables which are mean zero and independent. For the variance

$$\begin{aligned}&\mathrm{Var}(\alpha_{ij}) \\ \leq\ & E\left[x_{ij}^2\left(I\left(\epsilon_i\left(t_k, u_l\right) \leq C\sqrt{q_n/n^5}||x_{Ai}|| + C||W(z_i)||\sqrt{d_n/n^5}\right)\right.\right. \\ & \left.\left. -I\left(\epsilon_i\left(t_k, u_l\right) \leq 0\right)\right)^2\right] \\ \leq\ & C\left(\sqrt{q_n/n^5}||x_{Ai}|| + ||W(z_i)||\sqrt{d_n/n^5}\right) \\ \leq\ & Cd_n n^{-3/2}.\end{aligned}$$

Then by Bernstein's inequality for some positive constant C

$$\begin{aligned} I_{nj2} &\leq N_1 N_2 \exp\left(-\frac{n^2\lambda^2/32}{Cd_n n^{-3/2} + Cn\lambda}\right) \\ &\leq N_1 N_2 \exp(-Cn\lambda) \\ &\leq Cn^{4q_n} n^{4q_n+4dJ_n} \exp(-Cn\lambda) \\ &\leq C\exp((8q_n + 4dJ_n)\log(n) - Cn\lambda) \end{aligned}$$

Therefore using assumptions $\log(p_n) = o(n\lambda)$, $n^{-1/2}q_n = o(\lambda)$ and $n^{-1/2}dJ_n = o(\lambda)$ then

$$\begin{aligned} \sum_{j=q_n+1}^{p_n} (I_{nj1} + I_{nj2}) &\leq Cp_n \exp((8q_n + 4dJ_n)\log(n) - Cn\lambda) \\ &\leq C\exp(\log(p_n) + (8q_n + 4dJ_n)\log(n) - Cn\lambda) = o(1). \end{aligned}$$

□